

NATIONAL LIBRARY OF MEDICINE, NIH
BOARD OF SCIENTIFIC COUNSELORS
MEETING MINUTES
October 15 & 16, 2020

The Board of Scientific Counselors of the National Library of Medicine (NLM) convened by webcast on October 15, 2020, between 11:00 a.m. and 4:45 p.m. and on October 16, 2020, between 11 a.m. and 3:15 p.m. The meeting was open for viewing via NIH VideoCast.

BSC Members Participating

Kevin Johnson, MD, Vanderbilt University Medical Center (*BSC Chair*)
Hyun Min Kang, PhD, University of Michigan
Kateryna Makova, PhD, Penn State University
Susan Matney, PhD, Intermountain Healthcare (participated Oct. 15 only)
Ming Jack Po, MD, PhD, Google Health
Lucila Ohno-Machado, MD, PhD, University of California
Katherine Pollard, PhD, University of California
Steven Salzberg, PhD, Johns Hopkins University
Donna Slonim, PhD, Tufts University
Pamela Soltis, PhD, University of Florida
Peter Tarczy-Hornoch, MD, University of Washington
Jessica Tenenbaum, PhD, Duke University

NIH Staff Participating

Olivier Bodenreider, MD, PhD, LHC, NLM
Patricia Flatley Brennan, RN, PhD, NLM
Milton Corn, MD, NLM (*BSC executive secretary*)
Charles Dearolf, PhD, OD, NIH
David Landsman, PhD, NCBI, NLM
Kim Pruitt, PhD, NCBI, NLM
Jerry Sheehan, NLM
Bart Trawick, PhD, NCBI, NLM
(additional staff watched by videocast)

NLM Investigators Receiving Reviews

Eugene Koonin, PhD, NCBI, NLM
Teresa Przytycka, PhD, NCBI, NLM
John Spouge, MD, PhD, NCBI, NLM

NLM Senior Scientist Receiving Review

Steve Sherry, PhD, NCBI, NLM

NLM Poster Presenters

Noam Auslander, PhD, NCBI, NLM

Ayal Gussow, PhD, NCBI, NLM
Jan Hoinka, PhD, NCBI, NLM
Soumitra Pal, PhD, NCBI, NLM

Nash Rochman, PhD, NCBI, NLM
Ariella Saslafsky, Postbaccalaureate, NCBI, NLM

Day 1: Thursday, October 15, 2020

1. Welcome, Introductions, Scheduling – Kevin Johnson

Dr. Johnson welcomed participants to the meeting and introduced the three new BSC members: Drs. Hyun Min Kang, Lucila Ohno-Machado, and Peter Tarczy-Hornoch.

2. Remarks from NLM Director – Patricia Flatley Brennan

Dr. Brennan thanked the BSC members for their service and the NLM intramural investigators for their work. She briefly updated the BSC on several activities, including the continuing telework status of staff, NLM's ongoing renovations in the main library and the Lister Hill Center, launch of the Preprint Pilot for NIH-funded COVID-related publications, and collaboration with NIH on data access and sustainability. Dr. Brennan also described three NIH projects that NLM is participating in: UNITE, a project to address the health effects of systemic racism; Science of Senescence, which is examining cell aging; and All of Us, which is aiming to create a database with health information on one million people. In addition, Dr. Brennan updated the BSC on NLM's research program, noting that NLM will begin the search for a Scientific Director in the spring of 2021 and that a new training director will be starting in October 2020.

Dr. Brennan highlighted three areas in which NLM is particularly interested in BSC feedback: how to make the best use of BSC closed sessions with investigators and ensure that key issues requiring administrative engagement are communicated to the Scientific Director and NLM leadership; emerging ideas and investments NLM should be considering as it updates its Strategic Plan; and the role of the BSC play in evaluating NLM programs of research.

Discussion

The BSC discussed the synergy between NLM's research and its services and the role the BSC could play in advising NLM about research challenges that need to be met in order to create new services.

3. Presentation and Review of Eugene Koonin, Senior Investigator

Dr. Koonin focused his presentation on one area of research in his group: evolution and taxonomy of viruses. He described his group's work in "metaviromics," or viral metagenomics, for discovery of new viruses, as well as the group's research into the evolution of virus genomes, which resulted in a new virus taxonomy that has been adopted by the International Committee on Taxonomy of viruses.

Dr. Koonin also briefly mentioned seven other principal areas in which his group conducts research:

- Evolution and origin of human pathogenic viruses, including coronaviruses
- Classification, functions and evolution of bacterial and archaeal antiviral defense systems, in particular CRISPR-Cas adaptive immunity
- Mining genomic and metagenomic databases for novel viruses, mobile genetic elements and defense systems
- Evolution in tumors and evolution of cancer genes
- Mathematical and physical models of evolutionary processes informed by comparative genomics; theory of microbial genome evolution
- Evolution of biological complexity; general theory of genome evolution
- Collaborative research on functional characterization of biologically important cellular systems and individual proteins

Dr. Koonin concluded his presentation with a discussion of upcoming areas of research for his group, including more metaviromics and single-cell viromics, discovery of new virus classes and orders, improving family/general-level classifications through further data and extensive evolutionary analysis, looking at origin and taxonomy of orphan viruses, examining the origin of “mysterious” viruses such as those of hyperthermophilic archaea, and deep reconstructions of certain viromes.

Following Dr. Koonin’s presentation there was a brief Q&A, after which the BSC went into closed session with him.

4. Presentation and Review of Teresa Przytycka, Senior Investigator

Dr. Przytycka noted that her research group works on developing novel algorithms to address biological questions in network biology, gene regulation, and analysis of new types of experimental data. The computational questions the group addresses often emerge from collaborations with experimental groups.

In her presentation to the BSC, Dr. Przytycka focused on her group’s work on “systems biology dissection of the mutational landscape in cancer.” She described two complementary perspectives in looking at cancer mutations – the network-centric view and the data-centric – and described her group’s work in these areas, including development of the BeWith method for identifying a set of gene modules that are enriched within the module and between modules, and the NETPHIX (NETwork-to-Phenotype association with eXclusivity) method for predicting complementary drug sensitivity in cancer. She then presented her group’s method for combining network-based methods with data-centric views to uncover etiologies of mutational patterns in cancer.

Dr. Przytycka briefly touched on her group’s other research, including:

- Signature Estimation – statistical analysis of mutational signatures
- SigMA – Hidden Markov models for local dependences between mutations
- NetREX – method for construction of context-specific Gene Regulatory Networks
- Co-SELECT – dependence of TF binding on DNA shape

- scPopCorn – method for comparative analysis of single-cell experiments
- AptaSuite – set of novel algorithms and professional-level software for analysis of HT-Selex data
- AptaBlocks – method for RNA-based drug design

Dr. Przytycka also noted her group's ongoing and future research. These projects include:

- Use of mutational signatures to estimate the relation of smoking and biological processes relevant to COVID-19
- Analysis of the relation between non-B-DNA structures, cancer type, and mutational signatures
- Deconvolution of the primary and secondary causes in mutational signatures
- Construction of cell-type specific Gene Regulatory Network from noisy and incomplete data
- Study of the evolution of tissue-specific gene expression

Following Dr. Przytycka's presentation there was a brief Q&A, after which the BSC went into closed session with her.

5. Presentation and Review of John Spouge, Senior Investigator

Dr. Spouge began his presentation with a brief description of his background and his work history. He noted that he has always had a small group and currently is not supervising anyone, following the departure of a staff scientist in 2018 and a post-doctoral fellow in 2019. Two summer students were to join his group but ultimately did not because of COVID-19.

Dr. Spouge described several projects that he worked on during the 4-year period being reviewed:

- The genetics of HIV or other pathogens immediately after infection of a single host – Work in this area is continuing.
- A rapid algorithm for predicting the alternative structures of RNA switches – This project resulted in development of a conditional probability (CP) algorithm for predicting alternative RNA structures that was more than 1000 times faster than the sampling-clustering (SC) method to which CP was compared.
- Barcoding and exact k-mer matching for taxonomic identification of unknown biological samples – Dr. Spouge is collaborating with NCBI's Information Engineering Branch to develop NCBI software using diagnostic k-mers for taxonomic identification.
- The conservation of sequence pattern and genomic position due to functional constraints – This project was completed following publication of results.
- A statistical test for identifying domain alignment positions associated with disease-causing mutations – This project, which involved development of an alternative statistical test for researchers examining the domain mapping of disease mutations, was completed.
- A statistical test and efficient algorithm for finding overrepresented herpesvirus microRNA motifs on circular RNAs – This completed project involved developing a requested statistical test.

- COVID-19 – This project will estimate the parameters controlling the initial exponential growth and age structure of COVID-19.

Following Dr. Spouge’s presentation there was a brief Q&A, after which the BSC went into closed session with him.

6. Poster Session

Six NLM researchers who work with Drs. Koonin or Przytycka presented their posters:

- Noam Auslander, PhD, Koonin group – Insights into cancer progression from analysis of tumor mutational landscapes with machine learning approaches
- Ayal Gussow, PhD, Koonin group – Genomic determinants of pathogenicity in SARS-CoV-2 and human coronaviruses
- Nash Rochman, PhD, Koonin group – Deep phylogeny of cancer drivers and compensatory mutations
- Jan Hoinka, PhD, Przytycka group – AptaSuite - An in-silico framework for the analysis of HT-SELEX experiments
- Soumitra Pal, PhD, Przytycka group – EvoGeneX: stochastic modeling of gene expression evolution
- Ariella Saslafsky, Postbaccalaureate, Przytycka group – Learning patient history from mutational signatures with applications to COVID-19

7. BSC Discussion (Closed session)

Day 2: Friday, October 16, 2020

8. Presentation and Review of Steve Sherry, Senior Scientist

Dr. Sherry began his presentation by noting that while he currently is serving as NCBI Acting Director, for much of the review period he had the dual roles of Deputy Director of the Information Engineering Branch (IEB) Data Services Division and head of IEB’s Sequence Enhancements Program, where his work primarily focused on developing new infrastructure and services to advance federated data science.

Dr. Sherry focused his presentation on the work NCBI is doing to move its resources into the cloud. He described how, as part of NIH’s STRIDES initiative, NCBI moved the Sequence Read Archive – one of NIH’s largest and most diverse datasets, representing genome diversity throughout the tree of life – onto Amazon and Google cloud platforms. Dr. Sherry explained the benefits to both researchers and NIH of the cloud. For researchers, advantages include rapid access to large data sets, sharing data from a central location, elastic compute power, and reproducible analytical processes. For NIH, advantages include the ability to combine datasets across institutes for maximum statistical power, usage analytics, and transparent cost accounting.

Dr. Sherry described the Sequence Data Delivery Pilot (SDDP) project, which delivered a data sharing framework that enabled IC-funded cloud storage of large datasets. SDDP enabled cloud-based analysis of datasets by authorized users without creating additional copies of datasets, and it introduced a model of funding users for cloud-based compute costs. SDDP leveraged established NIH procedures for authorizing user access to potentially identifiable human genome data in a manner consistent with dbGaP access approvals.

Dr. Sherry also described work to develop a federated login. The framework incorporates traditional identity and access management (IAM) principles and utilizes token-based security for distinguishing between identity and access privileges.

Following Dr. Sherry's presentation there was a brief Q&A, after which the BSC went into closed session with him.

9. BSC Discussion (Closed session)

10. Report to Acting Scientific Director – Kevin Johnson

The BSC met in closed session with Dr. Charles Dearolf, from NIH's Office of Intramural Research, and Dr. Milton Corn, NLM Acting Scientific Director.

11. IEB Overview – Kim Pruitt and Bart Trawick

Drs. Pruitt and Trawick presented highlights of accomplishments at NCBI's Information Engineering Branch (IEB) over the last year, including COVID-19/SARS-CoV-2 activities. Highlights they presented included:

- NCBI data is used extensively, and NCBI often ranks as the most visited government website, according to analytics.usa.gov statistics (e.g., 33 million visits over 7 days for NCBI & 17 million for PubMed)
- Accomplishments towards the goal of moving resources to the cloud include: making 40 terabytes of SRA data available on Amazon and Google clouds as part of NIH's STRIDES initiative; moving PubMed website and database to cloud; beginning movement of ClinicalTrials.gov to cloud; and moving or starting to move several tools to the cloud, including the prokaryotic genome annotation pipeline (PGAPx), BLAST databases & CGIs, the read assembly & annotation pipeline tool (RAPT), and Elastic-BLAST
- NCBI has a proposal for a cloud-based Research Organism Ecosystem that will build upon its existing infrastructure and provide a central portal for all sequences as well as shared tools and will enable scalable analysis
- The new cloud-based PubMed was released in May 2020 and, among other enhancements, has made the mobile experience much better for users, as demonstrated by significantly improved net promoter scores (66.7 in 2020 vs. 37.4 in 2019)

- ClinicalTrials.gov is on path to move to the cloud, both for the website and submissions, and has been engaging extensively with stakeholders for input on the modernization
- In March, IEB released a major new dataset called Allele Frequency Aggregator (ALFA), which provides population frequency data for 447 million variants and 12 different ancestry populations; the open-access dataset, which is an aggregate of dbGaP data, is being presented through dbSNP and will soon be available through NCBI's ClinVar resource
- Upon receipt of the first SARS-CoV-2 sequence in January, IEB quickly moved to facilitate submissions and rapid access to the sequences
- IEB created a SARS-CoV-2 Resource page that provides easy access to relevant data/tools, including SARS-CoV-2 sequences in GenBank and SRA, the NCBI Virus SARS-CoV-2 Data Hub, SARS-CoV-2-related compounds and substances, relevant genome expression studies in the GEO database, BLAST, related reference sequences, a sequence submission portal, COVID-19 clinical trials, and related literature (via PubMed, PubMed Central, and a curated COVID literature collection called LitCovid)
- PubMed Central (PMC) launched a COVID-19 initiative in response to a call from national science and technology advisors from a dozen countries for publishers to make their COVID-19-related publications and related data available in PMC; under the initiative, as of September, more than 50 publishers have made 95,000 articles accessible in PMC in formats that facilitate text mining and secondary analysis; the collection is being leveraged by the Allen Institute for AI to build the COVID-19 Open Research Dataset (CORD-19)
- In June, NIH launched a pilot to test the viability of making preprints resulting from NIH-funded research searchable through PMC; Phase 1 is focused on COVID-19-related preprints (>1,000 as of September)
- In the next few weeks PubMed expects to release a new Clinical Queries page to facilitate discovery of citations related to SARS-CoV-2 and COVID-19
- ClinicalTrials.gov provides access to >3,500 COVID-19-related studies and includes 2,500 studies from World Health Organization Clinical Trials Registry portal
- NCBI is participating in the CDC-led SPHERES (SARS-CoV-2 Sequencing for Public Health Emergency Response, Epidemiology and Surveillance) initiative
- IEB has streamlined the submission process for SARS-CoV-2 sequences, with a new submission wizard and an automated processing pipeline that allows for release of sequences within an hour
- The NCBI Virus SARS-CoV-2 Resource provides easy search and retrieval of subsets of data, such as sequences from Wisconsin, and provides analysis tools and graphical widgets
- NCBI's Datasets Download resource makes it easier to download genomic datasets (e.g., nucleotide sequences and corresponding protein sequences) and is in the process of being connected to NCBI's Virus Hub
- NCBI SRA SARS-CoV-2 datasets are being made freely available on Amazon Web Services cloud, with no egress (download) charges
- IEB is developing several SRA SARS-CoV-2 detection and analysis tools (e.g., to detect SRA samples containing SARS-CoV-2, to build *de novo* assemblies, and to calculate nucleotide variants)

12. LHC Reorganization Update – Olivier Bodenreider

Dr. Bodenreider outlined the reorganization plan for Lister Hill Center (LHC), noting that the plan is still being reviewed by NIH, and thus some details might change. Among the reasons for the reorganization is to have greater focus on clinical data, to facilitate the transition to a single NLM intramural research program (IRP), and to centralize administrative and IT resources.

As part of the first phase of the reorganization, two of the six branches are no longer within LHC: the Office of High Performance Computing has been closed, and the activities of the Audiovisual Program Development Branch have moved elsewhere. Under the plan, the Office of the Director will remain, while the other three branches will be renamed. The new branches will be:

- Computational Health Research Branch – This branch will be a component of the NLM IRP and will consist of Principle Investigators and their groups. The branch will engage in research about clinical information processing, including the development of new advanced computational methods for clinical data.
- Applied Clinical Informatics Branch – This branch will not be a component of the IRP and will include staff scientists and others who are addressing broad problems in the use of clinical data, with an eye towards generalizable solutions that scale across institutional boundaries and applications. The branch will also help translate the insights discovered by the research branch into operational solutions.
- Scientific Computing Branch – Also not part of the IRP, this group will work on computing issues such as cloud resources.

Dr. Bodenreider noted that the planned structure parallels that of NCBI, with branches focusing on Research (NCBI's Computational Biology Branch), R&D (NCBI's Information Engineering Branch), and IT Services (NCBI's Information Resources Branch).

The focus on clinical data – including biomedical text and images, clinical analytics, and clinical standards – will guide future recruitment of PIs and staff. The research branch currently has three PIs and is recruiting for two others, one in clinical analytics and one in image processing. In addition, Dr. Michael Chiang, who was recently appointed as director of NIH's National Eye Institute, will have his research program in LHC's Computational Health Research Branch.

13. Plans for Future Meetings – Kevin Johnson

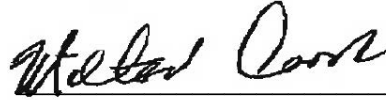
Dr. Landsman said he will identify potential dates for the next two BSC meetings and will circulate them to BSC members and NLM management. Dr. Johnson raised the issue of whether the BSC should commit to two virtual meetings or consider an in-person meeting in the fall of 2021. He said he would review comments from the written chat about the issue and that the group could discuss further via email.

Before closing the meeting, the BSC approved the minutes from their spring 2020 meeting.

A handwritten signature in black ink, appearing to read "Kevin Johnson", written over a light gray rectangular background.

11/20/20

Dr. Kevin Johnson, Chair
Board of Scientific Counselors

A handwritten signature in black ink, appearing to read "Milton Corn", written over a light gray rectangular background.

11/20/2020

Dr. Milton Corn (Date)
Acting Scientific Director, NLM