

**NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION  
NATIONAL LIBRARY OF MEDICINE, NIH**

**BOARD OF SCIENTIFIC COUNSELORS  
MEETING MINUTES  
April 9, 2019  
9:00 a.m. – 3:00 p.m.**

The Board of Scientific Counselors of the National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), convened on April 9, 2019 in the NLM Board Room, Bethesda, Maryland. The meeting was open to the public.

**NCBI BSC Members Present**

David Relman, M.D., Stanford University (*BSC Chair*)  
Kateryna Makova, Ph.D., Penn State University  
Katherine Pollard, Ph.D., University of California – via teleconference/webex  
Donna Slonim, Ph.D., Tufts University  
Pamela Soltis, Ph.D., University of Florida – via teleconference/webex  
Jianzhi Zhang, Ph.D., University of Michigan  
James Ostell, Ph.D., NCBI, NLM (*BSC Executive Secretary*)

**Lister Hill BSC Member Present**

Kevin Johnson, M.D., Vanderbilt University Medical Center – via teleconference/webex

**NLM Staff Present**

Richa Agarwala, Ph.D., NCBI, NLM  
Dennis Benson, Ph.D., NCBI, NLM  
Patricia Flatley Brennan, Ph.D., NLM  
Janet Coleman, NCBI, NLM  
David Landsman, Ph.D., NCBI, NLM  
Kim Pruitt, Ph.D., NCBI, NLM  
Valerie Schneider, Ph.D., NCBI, NLM  
Greg Schuler, Ph.D., NCBI, NLM  
Jerry Sheehan, NLM  
Steve Sherry, Ph.D., NCBI, NLM  
Bart Trawick, Ph.D., NCBI, NLM

**Others Present**

Charles Dearolf, M.D., OD, NIH  
Hannah Valentine, M.D., MRCP, NIH

**I. Welcome and Introductions**

Dr. Relman called the meeting to order at 9:00 a.m. Meeting participants introduced themselves.

Dr. Brennan thanked the BSC members for their service, in particular noting that Kevin Johnson, from the Lister Hill BSC, was attending his second NLM BSC meeting of the week.

## **II. NLM Director's Update – Patricia Flatley Brennan, Ph.D.**

Dr. Brennan updated the BSC on activities and new directions related to NLM's Strategic Plan, which was approved approximately 18 months ago. She structured her presentation on the three pillars of the Strategic Plan:

### **Accelerate discovery and advance health through data-driven research**

Dr. Brennan said NLM has a vision of the future where it fosters an ecosphere of discovery, enabled by an integrated research resource where the literature is well connected to clinical trial information, participant data, pathways, protocols, models, and other information. Each of these areas requires a foundation of structure, metadata, and ways of coding, tagging, and building grammars. She noted that NLM recognizes that there is increasing emphasis on building methods and documenting these methods to make them reusable.

Interconnecting the areas mentioned above will require expansion of NLM's research operation, as recommended in the Blue Ribbon Panel Review of NLM. In conjunction with those recommendations, NLM is in the process of recruiting three new investigators to its intramural research program. Top candidates from a first round of interviews have undergone a second set of interviews, and the new candidates are expected to be selected and in place by the fall of 2019. A second set of three investigators may be brought on board in 2020. NLM also is in the process of recruiting a scientific director to oversee NLM's two intramural research units (in NCBI and in the Lister Hill Center). Dr. Brennan explained that the intent is to maintain the integrity and identity of the two divisions, but to unify them under a single scientific director in an effort to accelerate research, including methods development and the use of new tools.

Other organizational changes at NLM include a new Chief Health Data Standards Officer (in the Office of the Director) and the elimination of the Specialized Information Services (SIS) division; SIS staff and responsibilities have been integrated within other NLM divisions.

NLM also will be improving its physical infrastructure, including unifying the ClinicalTrials.gov staff within a single area, creating more efficient administrative spaces, and improving the reliability of the data center. Planned improvements to the technological infrastructure include modernizing the ClinicalTrials.gov software platform and improving dbGaP workflows. In addition, NLM has initiated an analysis of its portfolio of resources ("offerings") to better understand issues such as how the offerings serve the NLM mission, which services are interdependent on each other, and the cost of services.

### **Reach more people in more ways through enhanced dissemination and engagement**

Dr. Brennan noted that NLM has a broad range of resources and a broad range of users, from clinicians to Nobel Prize winners to citizen scientists to patients. NLM's primary service is to make the 21<sup>st</sup> Century Collection available, useful, and accessible to individuals. Dr. Brennan identified three key responsibilities: to serve as the custodian of data and information, to serve as a connector to data and information, and to assist in the discovery of data and information. NLM

currently is targeting three areas: improved relevance-based search results, which have been implemented in PubMed and PubMed Labs; automated annotation; and addressing personalized presentation and delivery of information.

Dr. Brennan described plans to renovate NLM's reading room, which was built many decades ago when it was used by a greater number of people. One of the challenges in renovating the space is that there are some restrictions because of its historical and architectural significance.

#### **Build a workforce for data-driven research and health**

In addition to efforts with NIH to improve data science training, NLM has started its own initiative, developed in conjunction with Booz Allen Hamilton, called "Data Science Readiness." In January, NLM began communicating to staff about plans for a survey and training. The data readiness survey was conducted about two weeks ago, and more than 700 staff participated (approximately half of all staff, federal and contract). The surveys provided a profile and indications of where staff could strengthen their skills. More than 100 staff have indicated interest in taking an intensive data science course that will be offered.

NLM also is a key part of NIH's Strategic Plan for Data Science, which provides a roadmap for modernizing the NIH-funded biomedical data science ecosystem. Dr. Brennan highlighted some of NLM's activities regarding data sharing, including PubMed Central's storage of publication-related supplemental materials and datasets that are up to 2GB in size. NLM can facilitate linkage to larger data sets in outside repositories. NLM also is extensively involved in NIH's STRIDES initiative for making data accessible in the cloud; Dr. Steve Sherry described that effort later in the BSC meeting.

Dr. Brennan closed her presentation with a suggestion that members visit the NLM exhibition on the politics of yellow fever in Alexander Hamilton's America.

#### **Q&A**

In response to a question about dbGaP, Dr. Brennan noted that in 2017 NIH put out a request for comments on the data submission, access and management processes for dbGaP in order to improve and streamline the processes. Extraction of data will be receiving significant attention from NLM, she noted. Dr. Kim Pruitt, Acting Chief of NCBI's Information Engineering Branch, added that there is an independent report on the dbGaP process flow, written results of which are expected soon.

### **III. Presentation and Review of Richa Agarwala, Ph.D., Senior Scientist, IEB**

Dr. Agarwala focused her presentation on three broad areas of work that she and her group engaged in over the last four years: assembly, whole genome multi-locus sequence typing (wgMLST), and exploratory areas (metagenomics). She noted that the majority of her time was spent optimizing the Pathogen Detection Pipeline (PDP) via projects related to assembly and wgMLST.

#### **Background on PDP**

NCBI is collaborating with FDA, CDC, USDA and others on a project aimed at more quickly resolving outbreaks of foodborne disease through rapid identification of pathogens. Participating agencies and public health labs submit whole genome sequence reads from food, the environment, and patients to NCBI, which analyzes the data using its PDP. The PDP produces reports that identify each isolate and place them in a species-wide context. The reports also identify SNPs that distinguish closely related isolates and provide trees that show the relationships within isolate clusters. These reports flow back directly to the original submitters and, in generalized form, to the public. Since 2013, the PDP has been used by public health officials to support more than 370 actions. Dr. Agarwala's role in PDP is to develop the algorithms and prototype software for testing and potential addition to the production pipeline.

### Assembly

Dr. Agarwala's group developed SKESA (strategic K-mer extension for scrupulous assemblies) software for producing de-novo assemblies for real-time pathogen detection by PDP. Previously, the PDP process included doing a reference-based assembly and at least one de-novo assembly and combining them; this process slowed production because of issues with the de-novo assemblers. Dr. Agarwala's group decided to develop a de-novo assembler that would be tailored to the PDP process. SKESA's heuristics, she explained, are designed to reduce the effect on the assembly of low-level contamination and strand-specific errors in Illumina sequencing (the dominant technology in the market). SKESA software, which has been made freely available through Github, has a very low error rate on Illumina reads and is much faster than the previous process. In addition to the PDP, SKESA is being used in NCBI's Sequence Read Archive (SRA) for 5 organisms: *Salmonella*, *E. coli* and *Shigella*, *Campylobacter*, *Listeria*, and *C. difficile*.

Dr. Agarwala's group also developed a hybrid assembler for certain genes of interest, such as antimicrobial resistance genes, which have repeats that are longer than the insert size, a situation that results in SKESA creating a contig break.

### wgMLST schemes

Dr. Agarwala described how her group developed software for generating wgMLST schemes. The schemes are necessary to produce clusters that use allele calls instead of bases for genome comparison, an approach that significantly reduces the footprint for comparison. The software has been used to produce schemes for 18 species of interest in the PDP.

To generate rapid reports (within an hour) to assist in investigations before SNP reports from the PDP are available, SRA does the assembly by SKESA, allele calling using the wgMLST schemes, and pairwise comparison using alleles for the five covered species. The reports include the five nearest neighbors and all neighbors with five or fewer allele differences.

Dr. Agarwala presented data showing how the new software has significantly reduced PDP processing time. With *Salmonella*, for example, in December 2017 it took 383 hours for 94,000 isolates; in March 2019, it took a fraction of the time (50 hours) for twice as many isolates (188,000).

### Exploratory (metagenomics)

Dr. Agarwala described her group's work related to a Mosaic bioinformatics community challenge on clinical strain detection. They used their SPRISM alignment software to detect clinical strains and species in metagenomic samples, a new application for the software. The group also released software, called Matchhits, that aligns RNAseq reads to genomes.

#### Q&A

Discussion following Dr. Agarwala's presentation included the practical advantages that have been realized through the PDP project. Dr. Agarwala explained that much of the feedback has not been quantitative, but rather more general information about how the tools helped resolve national food poisoning cases such as contaminated Blue Bell ice cream and caramel apples. In response to questions about the need to produce the reports within an hour, Dr. Agarwala noted that NCBI has been told that the reports are needed as soon as possible by CDC and FDA.

#### **IV. Presentation and Review of Greg Schuler, Ph.D., Senior Scientist, IEB**

Dr. Schuler, who currently leads a team called Creative Services that is within IEB's Customer Services Division, presented highlights of his work on a number of projects. Most of the projects he described involved improving usability of NCBI websites via features that enhanced searching, filtering and browsing capabilities.

#### MultiSensor

Sensors are software that attempts to predict what a user is seeking based on their query. If a sensor "fires," content related to the predicted search intent is displayed above the main search results. NCBI has been using sensors, such as the citation sensor, for a number of years. Because each sensor required a large development effort, Dr. Schuler undertook development of MultiSensor, code that covers a wide range of query types. Rules and dictionaries are stored in a separate repository from the code, and users of the software can supply completely different data. MultiSensor has been very successful, with a 66% click-through rate, Dr. Schuler said.

#### Analytics: "trending articles," "articles frequently viewed together," & GenBank "hold until published" records

NCBI relies heavily on data from its internal tracking system (AppLog), and more recently products such as Google Analytics, to drive decisions about its websites, including design choices, such as font and colors, and new features. Dr. Schuler cited a recent project where he used an A/B test protocol to show that for one type of search result adding a context heading and increasing the size of the title link had a significant effect on the click-through rate. Dr. Schuler also noted his use of aggregate data to create new features, including a list of "trending articles" (articles where usage had been low and then suddenly increased – e.g., after Nobel prize announcements) on the PubMed homepage. He cited two other examples: 1) adding to the abstract page of some articles in PubMed a listing of articles frequently viewed together, which is one of the most highly used links on that section of the abstract page, and 2) using web analytics to create an automated process to identify GenBank records that should have been released upon publication of the associated article but were not.

#### Web design standards

Dr. Shuler began work on NCBI Web Design Standards (NWDS) in 2016 as NCBI was starting to migrate to a new Web development platform called Django. He based NWDS on the government's recently released U.S. Web Design Standards – which was meant to promote usable, accessible, and mobile-friendly websites across government – and added additional components such as a standard page header and footer. The footer includes JavaScript that reports which pages are using standards, allowing adoption of the standards to be measured. He estimated that about two dozen applications are now using the web standards. Dr. Schuler decided to use a “living style guide,” which incorporates the actual style sheets and user interface components that it is documenting.

### GQuery

GQuery, short for global query, handles the “all databases” search from the NCBI homepage and from within individual NCBI databases. The results page then shows hits in all of the 44 databases that are part of the Entrez search and retrieval system. Dr. Schuler assumed responsibility for GQuery in 2017 with 2 goals: 1) using it as a demonstration project for migrating from the old NCBI web platform (Portal) to the new platform (Django), and 2) enhancing the product over the long term to make it more useful. After the new platform was stable he began making gradual changes to fonts, colors and spacing to make the page conform to current NCBI design standards. MultiSensor and CitationSensor were then added, as well as a spellcheck feature that checked for misspelled words and incorrect accession numbers (a common mistake is to have too few or too many zeros).

### Knowledge and Information Search (KIS)

A Tier 1 project at NCBI for both FY18 and FY19, KIS aims to improve the user search experience, to create in essence a Google for scientists. The project includes multiple people from different teams at NCBI, including those who have expertise in user research, biological content, UX/UI design, software engineering, web analytics and customer communications. The project has introduced a number of new sensors to GQuery and several individual databases, including sensors for genes, genome assemblies, targeted loci, bacterial proteins, and antimicrobial genes. Dr. Schuler was primarily responsible for a new autosuggest feature, which shows as-you-type query suggestions. The user interface for the feature was based on open-source software, but the most important aspect was constructing underlying dictionaries that would include suggestions that would result in a display of sensor content. The dictionaries were compiled from relevant databases, with terms weighted based on data from search logs.

Dr. Schuler described the processes, known as “agile,” used in the KIS project, including “scrums” and “sprints” (a set of tasks that are set for the next two-week period). There is a 15-minute update meeting every morning where issues of the day are discussed and documented on a white board that is photographed and shared via Slack. Progress is tracked with a hand-drawn thermometer on the whiteboard. At the end of the two weeks there is a sprint review or a demo of what has been accomplished, he noted.

Dr. Schuler also described the process of usability testing and played the audio from a tape recording of a user testing session.

## **V. Unconscious Bias Training – Hannah Valantine, M.D., MRCP, NIH Chief Officer for Scientific Workforce Diversity**

Dr. Valantine's presentation centered on three areas: why diversity and inclusion matters, data on scientific workforce diversity, and mitigating implicit bias at NIH. She noted that the take-away message from her presentation is that improving diversity requires an integrated approach with many actions at the same time.

### Why diversity matters

Dr. Valantine noted that there are a lot of data suggesting that diversity in teams produces better science, a better ability to solve complex problems, and broadened scientific inquiry. For example, as more women have gotten into certain fields of science, such as cardiology, more questions are being asked of the research and new knowledge created.

### Scientific diversity data

Dr. Valantine presented a wealth of data relating to diversity in scientific fields. One chart showed that while biomedical Ph.Ds among underrepresented minorities (URM) grew 9-fold between 1980 and 2013, URM hiring for tenure-track assistant professor positions only grew about 2.5-fold; in contrast, well-represented (WR) populations showed a much more commensurate growth between Ph.Ds (a little more than 2-fold) and assistant professorships (a little under 2-fold). She showed similar data on female Ph.Ds: the number of female Ph.Ds began exceeding that of males around 2005, and yet males continued to be hired a much greater rate for tenure-track assistant professor positions throughout the 1980-2013 time period studied. Dr. Valantine cited these data in asserting that there is a pool of diverse candidates that can be tapped despite claims that it is difficult to find many diverse candidates.

At NIH, Dr. Valantine said, tenure track and tenured investigators are 72.8% male, which is similar to the national rates, and predominantly white (75.2%) and Asian/Pacific Islander (18.5%). African Americans account for 1.8%, Hispanics 4.4%, and Native Americans 0.2%. She also showed a breakdown by individual NIH institutes and centers.

### Implicit-bias

Dr. Valantine discussed the pervasiveness of implicit bias, citing the different words that are used to describe women versus men in letters of recommendation and studies showing, for example, that female concert musicians are selected at a much higher rate when the auditions are blinded. She noted that stereotypes begin early, citing a study where students were shown photographs of scientists and told to identify which people were scientists and which were teachers; the study found that the more feminine the person in the photograph, the more likely students were to guess they were a teacher rather than a scientist.

She reported results of a nationwide study published in 2012 that evaluated bias among 127 biology, chemistry and physics professors in evaluating application materials from undergraduate science students for a lab manager position. Both male and female faculty participants in the study rated the female students less competent and less suitable for hire; the female applicants also were offered a lower salary and less mentoring.

Dr. Valantine also cited a study in which 50% of medical students and residents were found to believe that black patients feel less pain than white patients and were more likely to suggest inappropriate treatments for the black patients.

Presentations such as this one can make a difference, Dr. Valantine said, pointing to results of a study she conducted that showed that after attending a similar presentation by department chairs, faculty members improved their scores on implicit bias tests. She indicated that the effect only appears to last a few months, so that periodic “boosters” are needed.

Dr. Valantine concluded with suggestions for “bias interrupters,” noting the importance of saying something at the time if there is a situation that might be explained by bias. One example is called “prove-it-again bias,” where some individuals are held to a higher standard due to stereotypes associated with their social identity. Dr. Valantine recommended being aware of such shifting standards and restating the evaluation criteria. She cautioned about situations where people say a candidate is not a good “fit,” which is often a code word for something else, and cited situations where a woman expresses an idea and a man who later repeats the same thing is given credit for the idea. In the later case it is useful to say something at the time, such as “yes, I remember Pam said that a few minutes ago.” Dr. Valantine also recommended that search committees have diverse membership.

## **VI. Report on the 2018 NLM Blue Ribbon Panel – Jerry Sheehan, Deputy Director, NLM**

Mr. Sheehan provided general background on the Blue Ribbon Panel review of NLM’s intramural research and training program, noting that such reviews of NIH Institutes and Centers (ICs) are typically done about every 10 years. He briefly described the Blue Ribbon Panel’s 11 recommendations and NLM’s activities to-date to implement those recommendations:

1. Work with NIH to significantly boost NLM’s investment in intramural research to support new independent investigators and embrace new opportunities for data science, informatics, and computational biology
  - Implementation: As noted by Dr. Brennan, NLM is actively recruiting new investigators for NLM’s intramural research program (IRP) and has increased its allocation of funding for investigators and associated staff.
2. Manage the IRP as one seamlessly connected, unified intramural research program with a single Scientific Director
  - Implementation: NLM is recruiting a Scientific Director to oversee its IRP and is creating a web page about the IRP.
3. Adopt one or more audacious, high-risk, high-reward projects to galvanize research across the organization and inspire the larger scientific community
  - Implementation: NLM believes the new Scientific Director should have the opportunity to identify these projects, although there has been discussion internally as well as input from NLM’s Board of Regents’ Research Frontiers subgroup.
4. Engage in a research portfolio evaluation and strategic plan to align NLM research priorities with the NLM Strategic Plan and research priorities of NIH and the broader biomedical research community



- **Implementation:** As with recommendation #3, NLM is awaiting the appointment of the new Scientific Director, but it has been looking at opportunities where the IRP could go beyond its current programs.
- 5. **Engage in a vigorous program of joint investigator appointments with other ICs to create a cadre of biomedical data science and informatics investigators with specific domain expertise who consider NLM to be their technical home**
  - **Implementation:** Mr. Sheehan noted that when this recommendation was presented to NIH leadership, they expressed concern about the practicalities of joint appointments. NLM is reviewing models used by other ICs where an investigator is on staff at one institute but is an adjunct faculty at another. NLM also is discussing the idea of NLM-sponsored symposia to convene relevant researchers from across NIH.
- 6. **Create mechanisms for identifying opportunities for moving research tools into services, based on current need and anticipated impact**
  - **Implementation:** This is an ongoing area of discussion. NLM is considering procedures for evaluating when a research project should transition into a service or move to another part of the organization. NLM has undertaken a portfolio analysis to understand its full set of offerings, including services and elements of the research program, that will be useful for this effort. Mr. Sheehan noted that NLM welcomes input from the NCBI and LHC BSCs to help it think through the process.
- 7. **Work with NIH to develop policies and procedures for optimizing successful recruitments given the competitive landscape for hiring in informatics and data science**
  - **Implementation:** NLM's current recruitments for investigators serve as test cases for its ability to attract top candidates to its IRP. NLM has engaged with the NIH Office of Scientific Workforce Diversity to improve recruitment of underrepresented populations.
- 8. **Restructure intramural research training into a single, unified training program with a designated Training Director**
  - **Implementation:** NLM has taken some steps forward. For example, NCBI staff participated in review of applicants for an LHC-announced program for research fellows, trainees, and visiting scientists. The next announcement will be NLM-wide. Additional steps are expected following recruitment of a new Scientific Director.
- 9. **Include broad metrics of scientific outcome, leadership, and impact, as well as publications and citations, in performance assessments of NLM's intramural research program and its researchers**
  - **Implementation:** NLM plans to seek input from the BSCs on this recommendation.
- 10. **Convene a single Board of Scientific Counselors with sufficient scientific breadth and expertise to evaluate the full set of activity within a unified intramural research program**
  - **Implementation:** NLM has begun the administrative process of establishing a single NLM BSC. The first meeting of the new BSC is tentatively planned for November 12-13. Initial membership will be the members of the existing NCBI and LHC BSCs.

11. Work with NIH to renovate and redesign intramural research labs and shared spaces to promote greater collaboration among research groups, optimize collective use of research equipment, and enable 21<sup>st</sup>-century team science
  - Implementation: NLM has begun planning to renovate space for the new investigators and staff and is developing a longer-term plan for renovating Building 38.

## VII. IEB Overview – Dr. Kim Pruitt

Dr. Pruitt abbreviated both of her presentations to the BSC because the meeting was running behind schedule. In this first presentation she briefly described each of the Information Engineering Branch's 16 "offerings" that emerged from NLM's portfolio analysis.

PubMed – PubMed is an essential, free citation resource that connects researchers, clinicians, healthcare providers, and the general public to biomedical literature and data. It contains almost 30 million records and has almost 52 million users/month.

PubMed Central – PMC provides free access to the full text of publicly funded research across scientific disciplines. It has 5.4 million records and 49.7 million users/month. More than 1 million papers have been deposited under the NIH Public Access Policy. The addition of an "associated data box" in November 2018 resulted in a 30% increase in daily downloads of supplementary material.

PubChem – PubChem helps researchers make sense of the biological roles and health effects of chemicals on human health and the environment. Some of NLM's TOXNET databases are being incorporated into PubChem as part of the consolidation of NLM's SIS division into other parts of NLM. PubChem contains about 350 million records and receives 4 million users/month.

ClinicalTrials.gov – ClinicalTrials.gov is recognized as the world's pre-eminent database for clinical trial registration and results reporting and serves as a permanent archive of clinical trials and other types of clinical research. NLM is requesting funding from NIH for modernization of ClinicalTrials.gov. The database contains more than 300,000 records and is accessed by an average of 116,000 visitors each day.

dbGaP – The database of Genotypes and Phenotypes stores and disseminates human genotype and associated phenotype data; a portion of the data is available only through controlled access. dbGaP contains data on about 1,280 studies and receives 27,100 users/month. dbGaP was introduced more than 10 years ago and NLM is hoping to modernize it in the coming year.

dbSNP – dbSNP is an authoritative, open-access and permanent catalog of short human genetic variations. This year a new notation and API service was put into operation. The architecture of the database was modernized to allow for projected growth to 1 billion variations in 2019 (dbSNP has 1.8 billion submitted variations but 677 million records; the SNP pipeline was redeveloped to condense individual submissions into a non-redundant set, which is much more effective and efficient).

**GTR and MedGen** – The Genetic Testing Registry includes molecular, cytogenetic and biochemical clinical and research tests. It contains data from over 500 labs on more than 55,000 tests for 11,000 conditions and 16,000 genes. MedGen organizes information related to human medical genetics, such as attributes of conditions with a genetic contribution.

**ClinVar** – Built on top of dbSNP, ClinVar contains assertions about the association between phenotypes and genetic variations. The database has been extremely valuable in offering transparency when there are conflicting assertions about such associations. ClinVar contains over 500,000 records and receives more than 93,000 users/month.

**GenBank** – GenBank is NLM’s foundational database of genome sequences, with 341 million records and more than 600,000 users/month. NCBI continues to automate GenBank processing to be able to handle more data without growing staff. For example, more automated submission and processing pipelines recently were put in place for Noravirus and dengue virus.

**SRA** – The Sequence Read Archive is the largest publicly available repository of next-generation sequence data, with more than 7.5 million records and 78,600 users/month. On a monthly basis SRA delivers almost 690 terabytes of data. SRA’s movement to a cloud environment was addressed later during the meeting by Dr. Steve Sherry.

**GEO** – The Gene Expression Omnibus database is the largest collection of richly annotated, open-access gene expression and epigenomics datasets from all branches of life. It contains 131 million records and has more than 100,000 users/month.

**RefSeq** – This database of reference sequences is a comprehensive collection of integrated, annotated sequence and gene information that is built on top of GenBank and SRA. RefSeq genomes have been annotated using a single method as opposed to the variety of methods one would find if looking across GenBank. NCBI has been working with EMBL-EBI in Europe to improve and harmonize human annotation for clinical and basic research. RefSeq contains 192 million records and receives more than 1 million users/month.

**Pathogens** – The Pathogen Detection Project integrates data on genomic sequences from bacterial pathogens originating in food, environmental sources, and patients. It quickly clusters and identifies related sequences to uncover potential food contamination sources, helping public health scientists investigate foodborne disease outbreaks. FDA has used results from the Pathogen Detection pipeline in 370 regulatory and compliance actions. In the last year, 90,000 isolates were analyzed, and the resource had 1,890 interactive monthly users.

**CDD/Sparcle** – CDD/Sparcle is a frequently updated and curated sequence analysis resource for protein sequences, with over 190,000 records and about 113,000 monthly users. It offers a representation of protein domains and domain architectures.

**SeqTools/BLAST** –NCBI SeqTools, such as the flagship BLAST collection of sequence comparison tools, enable researchers to efficiently and effectively access and use NCBI offerings. For example, the Genome Workbench tools is used by the Genome Reference Consortium for curation of the human reference genome assembly and by GenBank for

submission processing. Other tools include the Genome Data Viewer, the Sequence Viewer and analysis resources for users with more targeted research interests, such as Virus Variation.

Following Dr. Pruitt's presentation, one of the BSC members asked whether there was a formal process for evaluating the new features and the presentation of resources such as dbGaP. Dr. Pruitt explained that NCBI does have a formal process which includes user surveys, usability tests, a/b testing on the NCBI Labs website, and web analytics to understand user interaction with websites.

## **VIII. IEB FY2019 Tier 1 Initiatives**

Dr. Pruitt described NCBI's Tier 1 projects for fiscal year (FY) 2019. She noted that one of the lessons from the five Tier 1 projects in FY18 was that five was too many, and therefore only three were selected for FY19, although one of those projects – NIH STRIDES – has turned out to be more like four or five projects. Of the three Tier 1 projects, NIH STRIDES is the highest priority, followed by PubMed 2.0, and then Knowledge and Item Search (KIS).

### **NIH STRIDES**

Dr. Pruitt briefly described the STRIDES initiative because the next presentation, by Dr. Steve Sherry, provided a detailed explanation of this project. She noted that STRIDES is NCBI's highest-level Tier 1 project because of its importance to NIH and its large impact on the use of cloud resources across the scientific community.

### **PubMed 2.0**

This was a Tier 1 project in FY18 as well as FY19. The aim is to transform PubMed into a modern hub with a fast, reliable and intuitive search that connects people to the world's leading sources of biomedical information. Goals for FY19 include implementing a modern cloud architecture and a high-quality search, and making sure it is a fast and reliable experience that is appreciated by users. Goals include providing the service with 99.9% availability, a 15% improvement in user satisfaction based on user surveys, and 25% of users voluntarily and preferentially using the new PubMed.

Dr. Pruitt presented a slide showing the timeline for delivery of various PubMed functions. "Save" and "share" functions were introduced in the first quarter, and advanced search in the second quarter. Currently work is proceeding on modernization of MyNCBI functions. Work on LinkOut, API support, and tool integration are planned for the remainder of the year.

A preview of the new PubMed can be seen on PubMed Labs, NCBI's site for experimenting with new PubMed designs and functionality. Despite limited advertising of PubMed Labs, an increased number of users have been returning to the site and using it because of the improved user interface. Dr. Pruitt noted that when the new PubMed is introduced users will be able to go back to the old site during a transition period.

### **Knowledge and Item Search (KIS)**

KIS also was a Tier 1 project in FY18. The goal of the project at that time was to deliver expected, high-value results for text searches where the users have known sequence items in mind, regardless of the queried database. A number of sensors were developed to help users find their results, and additional features were implemented such as auto-suggest. In FY19 the goal is being expanded to build “Google for scientists” – a better NCBI search experience so that biomedical researchers, data scientists and others can find the information they need. Key challenges for users include that data at NCBI is siloed and users don’t necessarily know which database to use; the number of search results can be overwhelming; users don’t know which is the “best” result; and curation-intensive datasets may not be easily discoverable.

Dr. Pruitt described a few examples of how search is being improved, including an effort to rescue searches in ClinVar where users get no results because of poorly formatted Human Genome Variation Society terminology. She explained how success is monitored through metrics such as user clicks and rescued searches as well as user feedback and whether or not 95% of page renders are delivered within one second. Future plans include evaluating development of APIs, work on protein family searches, improving search of protein names and variation, and incorporating datasets from SRA and GEO.

## **IX. SRA in the Cloud – Dr. Steve Sherry**

Dr. Sherry described NCBI’s involvement in NIH’s STRIDES (Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability) project, which aims to promote an ecosystem that advances genome research through cloud-centered compute environments.

### **Background**

Dr. Sherry noted that in addition to being ambitious in technical ways, the project has more general challenges in that a number of NIH ICs have built data silos that service their specific disease communities and grant activities. He said there are two sides to success: for the researcher, and for NIH. For the researcher, the value will be 1) rapid access to relevant data sets 2) powerful compute resources, and 3) platforms for collaboration. For NIH, the value will be in enabling data sets within and between ICs to be combined for maximum statistical power.

STRIDES is an NIH Office of the Director program that began as a way for NIH to obtain discounted cloud storage and compute services by negotiating on behalf of all the ICs. As part of this project, NCBI is putting the data from its Sequence Read Archive (SRA) in Google and Amazon clouds, under what is called the Sequence Data Delivery Pilot (SDDP).

### **SRA data migration schedule and details**

Starting in May, all new incoming data to SRA will be put on Amazon and Google cloud platforms in both the original format of the submission and the standard SRA format (ETL).

In July, the public SRA data (about 5 petabytes, consisting mostly of non-human data and the 1000 Genomes Project data) will be made available on the cloud platforms in ETL format.

By September 30, the controlled-access SRA data (also about 5 petabytes) will be made available. Availability of these data will be coordinated with the release of CIT's JWT software framework for controlling access to authorized users.

Lastly, the archive of SRA data that is in original format will be moved to the cloud platforms, with estimated completion in 2021. These data will not go on disk, but will be on slower backup platforms at Google and Amazon, as the expectation is that the data will be in less demand.

Similarly, as the data that is on disk become older and no longer used, it will be moved to the backup platforms. Users will be able to access any of these datasets in the backup storage by making a request; Dr. Sherry estimated that movement of the data to hot storage would only take about 24 hours and would be a modest expense for users.

### Study consents

Dr. Sherry described some of the issues around controlled-access data and study participant consents. Within the controlled-access portion of data in SRA and dbGaP, there are currently 815 different consent groups (e.g., in some studies participants only consented to have their data used for non-profit research relating to their disease). How to provide data access based on consents becomes complicated by the fact that some ongoing studies ask subjects their consent preference every six-months. Dr. Sherry explained how NCBI could use metadata to make sure that researchers are only obtaining data that matches current study consents.

### Controlling data access

Dr. Sherry explained the process that is envisioned for users to see a catalog of data that is available and the software to verify that they have the proper permissions and provide them with access. He noted that some of the details regarding user authorization are still being worked out with the NIH Center for Information Technology (CIT) and the Office of Data Science Strategy (ODSS).

### BLAST applications in the cloud

NCBI's Basic Local Alignment Search Tool (BLAST) for finding regions of similarity between sequences is a very popular tool that is hosted as a web service at NLM as well as available for standalone use that people can run locally. NCBI has developed two approaches for BLAST as data move to cloud environments: 1) to provide command-line BLAST+ that can run in Amazon and Google clouds (a prototype version of this has been developed and is expected to be released in June), and 2) to provide the NCBI Web BLAST service in the clouds, allowing the BLAST search to be executed there.

### Hackathons

Hackathons are one way to check that NCBI's cloud approach is readily accessible to users and appropriate for their work as well as to provide education to the community. The first cloud hackathon, in January, was on using metagenomic SRA datasets in the cloud to identify viral content. NCBI did contig assembly on metagenomic runs in order to reduce the number of objects that would be computed on. Among the goals was to assess whether the contig approach made sense: was it attractive to users, did it help reduce storage costs, did it help find viruses. About 40 people participated in the hackathon and separated into teams working on

different problems, addressing questions such as: “Do these unknown contigs have gene-like patterns?” “Are these really viruses?” “Do these contigs match known viruses in RefSeq?” “Do these unknown contigs match a structural domain?”

Three other cloud hackathons have taken place during the year and several more are planned, including ones on haplotype annotation, prokaryotic genome annotation, single cell transcriptomics, and virus hunting (part 2).

#### Data footprint

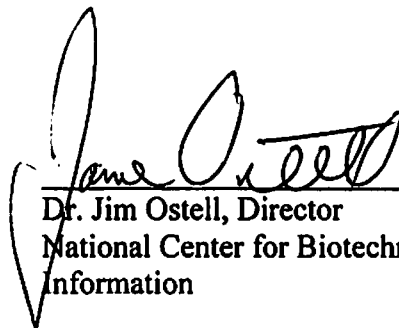
Dr. Sherry noted that the only way SRA can continue to support exponential growth in the size of data on a flat budget is to compress data. As sample depth is growing, with 50x coverage, there are other ways of revealing structure in the data. He noted that reducing quality scores from 8 bits to 2 would eliminate 70% of the data footprint, buying another year of storage. Also taking the independent reads and turning them into something longer like a contig (as was done in the virus hackathon) is another way to reduce the footprint.

#### **IX. Adjournment**

The BSC adjourned at approximately 3:00 p.m.

 7/21/2019

Dr. David Relman, Chair (Date)  
Board of Scientific Counselors

 7/25/2019

Dr. Jim Ostell, Director (Date)  
National Center for Biotechnology  
Information