EDUCATION

Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators

Lindsay Barone*[☯], **Jason Williams**[☯], **David Micklos**[☯]

DNA Learning Center, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, United States of America

☯ These authors contributed equally to this work.
* lbarone@cshl.edu

Abstract

In a 2016 survey of 704 National Science Foundation (NSF) Biological Sciences Directorate principal investigators (BIO PIs), nearly 90% indicated they are currently or will soon be analyzing large data sets. BIO PIs considered a range of computational needs important to their work, including high performance computing (HPC), bioinformatics support, multistep workflows, updated analysis software, and the ability to store, share, and publish data. Previous studies in the United States and Canada emphasized infrastructure needs. However, BIO PIs said the most pressing unmet needs are training in data integration, data management, and scaling analyses for HPC—acknowledging that data science skills will be required to build a deeper understanding of life. This portends a growing data knowledge gap in biology and challenges institutions and funding agencies to redouble their support for computational training in biology.

Author summary

Our computational needs assessment of 704 principal investigators (PIs) with grants from the National Science Foundation (NSF) Biological Sciences Directorate (BIO) confirmed that biology is awash with big data. Nearly 90% of BIO PIs said they are currently or will soon be analyzing large data sets. They considered a range of computational needs important to their work, including high performance computing (HPC), bioinformatics support, multistep workflows, updated analysis software, and the ability to store, share, and publish data. However, a majority of PIs—across bioinformatics and other disciplines, large and small research groups, and 4 NSF BIO programs—said their institutions are not meeting 9 of 13 needs. Training on integration of multiple data types (89%), on data management and metadata (78%), and on scaling analysis to cloud/HPC (71%) were the 3 greatest unmet needs. Hardware is not the problem; data storage and HPC ranked lowest on their list of unmet needs. The problem is the growing gap between the accumulation of big data—and researchers' knowledge about how to use it effectively.

OPEN ACCESS

Citation: Barone L, Williams J, Micklos D (2017) Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators. PLoS Comput Biol 13(10): e1005755. https://doi.org/10.1371/journal.pcbi.1005755

Editor: Francis Ouellette, Genome Quebec, CANADA

Published: October 19, 2017

Funding: This study is an Education, Outreach and Training (EOT) activity of CyVerse, an NSF-funded project to develop a "cyber universe" to support life sciences research (DBI-0735191 and DBI-1265383). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

This is a *PLOS Computational Biology* Education paper.

Introduction

Genotypic data based on DNA and RNA sequences have been the major driver of biology's evolution into a data science. The current Illumina HiSeq X sequencing platform can generate 900 billion nucleotides of raw DNA sequence in under 3 days—4 times the number of annotated nucleotides currently stored in GenBank, the United States "reference library" of DNA sequences [1, 2]. In the last decade, a 50,000-fold reduction in the cost of DNA sequencing [3] has led to an accumulation of 9.3 quadrillion (million billion) nucleotides of raw sequence data in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA). The amount of sequence in the SRA doubled on average every 6–8 months from 2007–2016 [4, 5]. It is estimated that by 2025, the storage of human genomes alone will require 2–40 exabytes [5] (an exabyte of storage would hold 100,000 times the printed materials of the U.S. Library of Congress [6]). Beyond genotypic data, big data are flooding biology from all quarters—phenotypic data from agricultural field trials, patient medical records, and clinical trials; image data from microscopy, medical scanning, and museum specimens; interaction data from biochemical, cellular, physiological, and ecological systems; as well as an influx of data from translational fields such as bioengineering, materials science, and biogeography.

A 2003 report of a National Science Foundation (NSF) blue-ribbon panel, headed by Daniel Atkins, popularized the term cyberinfrastructure to describe systems of data storage, software, high performance computing (HPC), and people who can solve scientific problems of the size and scope presented by big data [7]. The report was the impetus for several cyberinfrastructure projects in the biological sciences, including the NSF's CyVerse, the Department of Energy's KBase, and the European Grid Infrastructure and the European Life Sciences Infrastructure for Biological Information (ELIXIR) [8]. The Atkins Report described cyberinfrastructure as the means to harness the data revolution and to develop a "knowledge economy." Although people were acknowledged as active elements of cyberinfrastructure, few published studies have assessed how well their computational and cyberinfrastructure needs are being met.

In 2006, EDUCAUSE surveyed 328 information technology (IT) professionals, primarily chief information officers, at institutions in the US and Canada [9]. When asked about preferences for funding allocation, respondents rated training and consulting (20%) a distant second to infrastructure and storage (46%). This suggested that "training and consulting get short shrift when bumped against the realities of running an IT operation" [9]. Similarly, infrastructure and training emerged as important needs in a study done as part of the 2015 University of Illinois's "Year of Cyberinfrastructure" [10]. Faculty and graduate students responding to a survey ($n = 327$) said they needed better access to data storage (36%), data visualization (29%), and HPC (19%). Training was not addressed in the initial survey, suggesting that it was not viewed as integral to discussions of cyberinfrastructure. However, it emerged as a major need in follow-up focus groups ($n = 200$).

Over the last 4 years, CyVerse has taken the computational pulse of the biological sciences by surveying attendees at major professional meetings. Consistently and across different conference audiences, 94% of students, faculty, and researchers said that they currently use large data sets in their research or think they will in the near future ($n = 1{,}097$). Even so, 47% rated their bioinformatics skill level as "beginner," 35% rated themselves "intermediate," and 6% said they have never used bioinformatics tools; only 12% rate themselves "advanced" ($n = 608$); 58% felt their institutions do not provide all the computational resources needed for their research ($n = 1{,}024$). These studies suggest a scenario of big data inundating unprepared biologists.

Results

In summer 2016, we expanded upon our previous studies with a purposeful needs assessment of 704 principal investigators (PIs) with grants from the NSF Directorate of Biological Sciences (BIO). The respondents were relatively evenly dispersed among 4 major BIO divisions: Division of Biological Infrastructure (DBI), Division of Environmental Biology (DEB), Division of Integrative Organismal Systems (IOS), and Division of Molecular and Cellular Biosciences (MCB). These BIO PIs worked with a variety of data, with sequence, image, phenotype, and ecological data predominating (Fig 1). The vast majority (87%) said they are currently using big data sets in their research or will within the next 3 years. This is slightly lower than in our previous studies of meeting attendees, a large proportion of whom had a genomics focus or were students or early career researchers.

We asked BIO PIs to rate the importance of 13 computational needs in data analysis, data storage, sharing, and discovery and computational support and training. More than half of the PIs said that 11 of the 13 computational needs are currently important to their research. The proportions increased across all needs—to 82%–97%—when PIs considered what would be important 3 years in the future (Fig 2). Significantly more PIs who identified themselves as bioinformaticians said 9 of the current needs are important compared to PIs from all other disciplines. Significantly more PIs from larger research groups (greater than 5 people) said 7 of the current needs are important compared to those from smaller groups. Most of the differences between bioinformaticians and larger research groups persisted in their predictions of future needs (Table 1).

Significantly more PIs funded by DEB said 5 of the current needs are important compared to PIs funded through the other 3 NSF research divisions. However, differences between the 4 NSF divisions disappeared for predictions of future need, suggesting that computational needs will converge across all fields of biology in the future (Table 2).

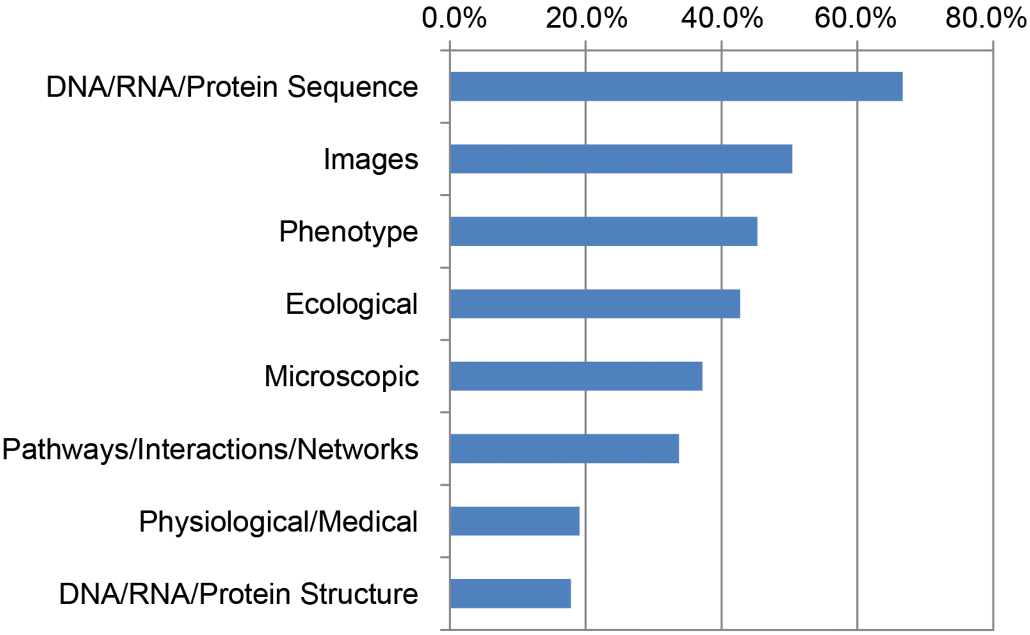

Fig 1. Major data types used by National Science Foundation (NSF) Biological Sciences Directorate (BIO) principal investigators (PIs).

https://doi.org/10.1371/journal.pcbi.1005755.g001

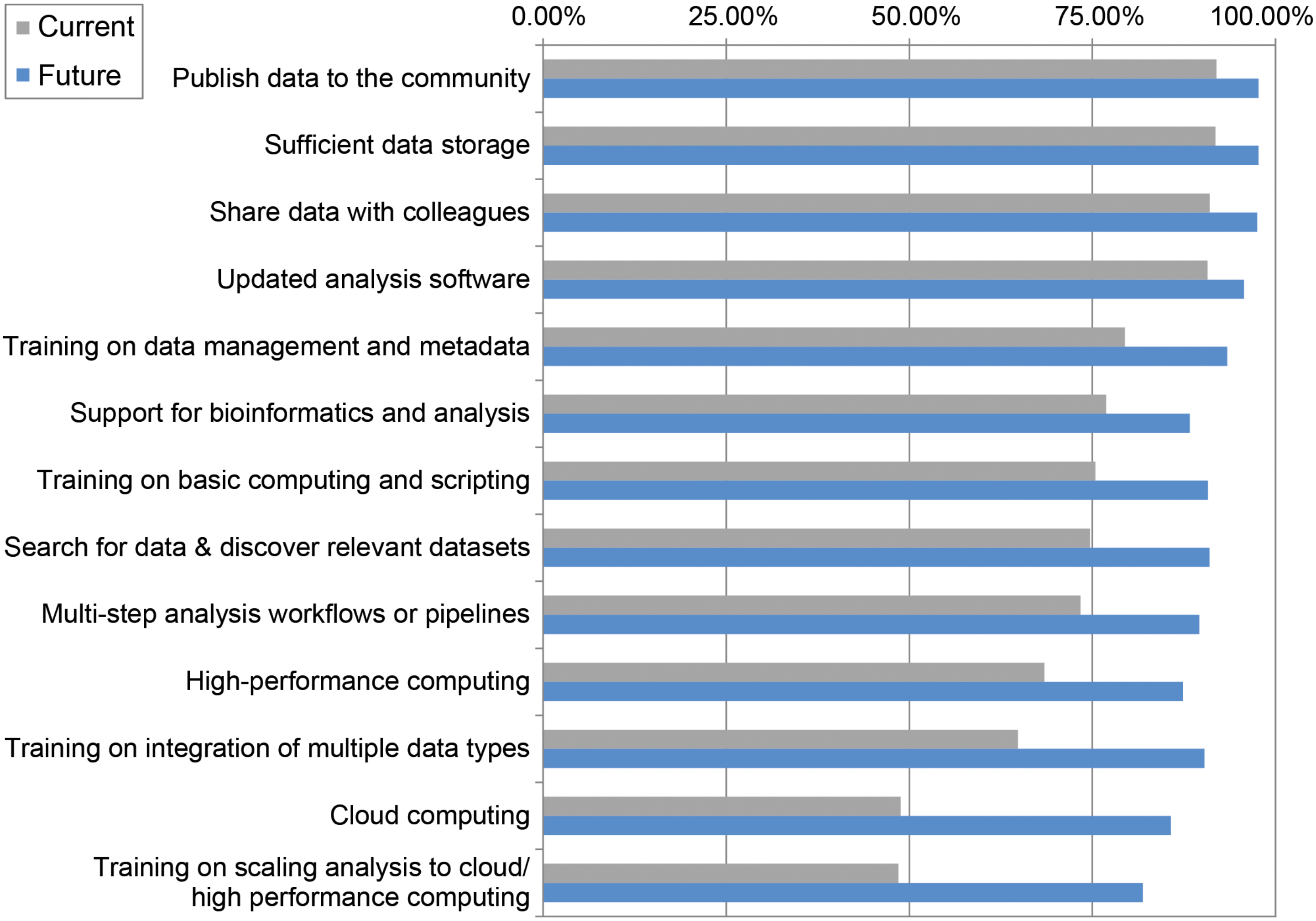

Fig 2. Current (grey) and future (blue) data analysis needs of National Science Foundation (NSF) Biological Sciences Directorate (BIO) principal investigators (PIs) (percent responding affirmatively, 387 ≤ *n* ≤ 551).

https://doi.org/10.1371/journal.pcbi.1005755.g002

A majority of PIs—across bioinformatics/other disciplines, larger/smaller groups, and the 4 NSF programs—said their institutions are not meeting 9 of 13 needs (Fig 3). Training on the integration of multiple data types (89%), on data management and metadata (78%), and on scaling analysis to cloud/HPC (71%) were the 3 greatest unmet needs. HPC was an unmet need for only 27% of PIs, with similar percentages across disciplines, different sized groups, and NSF programs.

Discussion

This study fills a gap in the published literature on the computational needs of biological science researchers. Respondents had all been awarded at least 1 peer-reviewed grant from NSF BIO and thus represent competitive researchers across a range of biological disciplines. Even so, a majority of this diverse group of successful biologists did not feel that their institutions are meeting their needs for tackling large data sets.

This study stands in stark contrast to previous studies that identified infrastructure and data storage as the most pressing computational needs [9, 10]. BIO PIs ranked availability of

Table 1. Current and future data analysis needs of National Science Foundation (NSF) Biological Sciences Directorate (BIO) principal investigators (PIs): Bioinformaticians versus others, large versus small research groups.

	Current needs				Needs in 3 years			
	Bioinformatics $151 \leq n \leq 153$	**All others** $91 \leq n \leq 399$	**Large group (>5 people)** $245 \leq n \leq 249$	**Small group (<5 people)** $293 \leq n \leq 298$	**Bioinformatics** $114 \leq n \leq 150$	**All others** $263 \leq n \leq 385$	**Large group (>5 people)** $187 \leq n \leq 246$	**Small group (<5 people)** $196 \leq n \leq 295$
Publish data to the community	0.97	0.90	0.94	0.90	0.97	0.98	**0.99**	**0.96**
Sufficient data storage	0.94	0.91	0.94	0.90	0.97	0.98	0.98	0.98
Share data with colleagues	0.93	0.90	**0.94**	**0.89**	0.97	0.98	0.98	0.97
Updated analysis software	0.93	0.90	0.92	0.90	0.96	0.96	0.96	0.95
Training on data management and metadata	**0.88**	**0.76**	0.83	0.77	**0.97**	**0.92**	**0.96**	**0.92**
Support for bioinformatics and analysis	**0.90**	**0.72**	0.80	0.75	**0.95**	**0.85**	0.90	0.87
Training on basic computing and scripting	**0.87**	**0.71**	**0.80**	**0.72**	0.94	0.90	0.92	0.90
Search for data and discover relevant data sets	**0.86**	**0.70**	0.74	0.75	0.96	0.89	0.91	0.91
Multistep analysis workflows or pipelines	**0.90**	**0.67**	**0.79**	**0.68**	**0.97**	**0.87**	**0.93**	**0.86**
High performance computing (HPC)	**0.89**	**0.61**	**0.75**	**0.63**	**0.96**	**0.84**	**0.91**	**0.84**
Training on integration of multiple data types	**0.78**	**0.60**	**0.76**	**0.56**	**0.95**	**0.88**	**0.94**	**0.88**
Cloud computing	**0.57**	**0.46**	**0.53**	**0.45**	**0.92**	**0.83**	**0.91**	**0.81**
Training on scaling analysis to cloud/ HPC	**0.71**	**0.40**	**0.57**	**0.41**	**0.94**	**0.76**	**0.86**	**0.78**

Percent responding affirmatively. Bold text indicates a statistically significant chi-square result between groups (bioinformaticians versus others; large research groups versus small).

https://doi.org/10.1371/journal.pcbi.1005755.t001

data storage and HPC lowest on their list of unmet needs. This provides strong evidence that the NSF and individual universities have succeeded in developing a broadly available infrastructure to support data-driven biology. Hardware is not the issue. The problem is the growing gap between the accumulation of many kinds of data—and researchers' knowledge about how to use them effectively. The biologists in this study see training as the most important factor limiting their ability to best use the big data generated by their research.

Closing this growing data knowledge gap in biology demands a concerted effort by individual biologists, by institutions, and by funding agencies. We need to be creative in scaling up computational training to reach large numbers of biologists at all phases of their education and careers and in measuring the impact of our educational investments. Metrics for a supercomputer are readily described in terms of petaflops and CPUs, and we can facilely measure training attendance and "satisfaction." However, answering unmet training needs will require a better understanding of how institutions are attempting to meet these needs—

Table 2. Current and future data analysis needs of National Science Foundation (NSF) Biological Sciences Directorate (BIO) principal investigators (PIs) by the NSF BIO division.

	Current needs				**Needs in 3 years**			
	Environmental biology 163 $\le n \le$ 168	**Molecular and cellular biosciences 85 $\le n \le$ 90**	**Biological infrastructure 116 $\le n \le$ 118**	**Integrative organismal systems 159 $\le n \le$ 161**	**Environmental biology 124 $\le n \le$ 162**	**Molecular and cellular biosciences 59 $\le n \le$ 85**	**Biological infrastructure 85 $\le n \le$ 117**	**Integrative organismal systems 108 $\le n \le$ 157**
Publish data to the community	**0.95**	**0.92**	**0.93**	**0.87**	0.99	0.98	0.96	0.98
Sufficient data storage	0.93	0.91	0.90	0.94	0.99	0.95	0.97	0.98
Share data with colleagues	0.95	0.90	0.91	0.88	0.98	0.99	0.96	0.97
Updated analysis software	0.92	0.88	0.91	0.91	0.95	0.93	0.95	0.99
Training on data management and metadata	**0.87**	**0.71**	**0.81**	**0.74**	0.94	0.91	0.95	0.92
Support for bioinformatics and analysis	0.80	0.83	0.72	0.76	0.89	0.90	0.88	0.87
Training on basic computing and scripting	**0.83**	**0.77**	**0.65**	**0.72**	0.94	0.89	0.85	0.92
Search for data and discover relevant data sets	0.75	0.77	0.77	0.71	0.93	0.93	0.91	0.88
Multistep analysis workflows or pipelines	**0.81**	**0.74**	**0.68**	**0.69**	0.93	0.88	0.92	0.86
High performance computing (HPC)	**0.77**	**0.68**	**0.64**	**0.63**	0.91	0.90	0.85	0.83
Training on integration of multiple data types	0.69	0.57	0.62	0.65	0.91	0.93	0.89	0.91
Cloud computing	0.56	0.41	0.50	0.46	0.87	0.85	0.84	0.87
Training on scaling analysis to cloud/HPC	0.55	0.46	0.50	0.41	0.86	0.78	0.79	0.80

Percent responding affirmatively. Bold text indicates a statistically significant chi-square result between BIO divisions.

https://doi.org/10.1371/journal.pcbi.1005755.t002

and how we can best assess their outcomes [11]. Some solutions already exist. For example, data sets available at the SRA provide almost unlimited entry points for course-based undergraduate research experiences (CUREs), which scale up discovery research in the context of for-credit courses. Participation in CUREs significantly improves student graduation rates and retention in science—effects that persist across racial and socioeconomic status [12, 13]. However, many biologists acquire skills for big data analysis on their own, in the midst of their careers. Software Carpentry and Data Carpentry [14] are volunteer-driven organizations that provide a cost-effective, disseminated model for reaching biologists outside of an academic classroom.

Reflected in the top 2 unmet needs of BIO PIs is the looming problem of integrating data from different kinds of experiments and computational platforms. This will be required for a

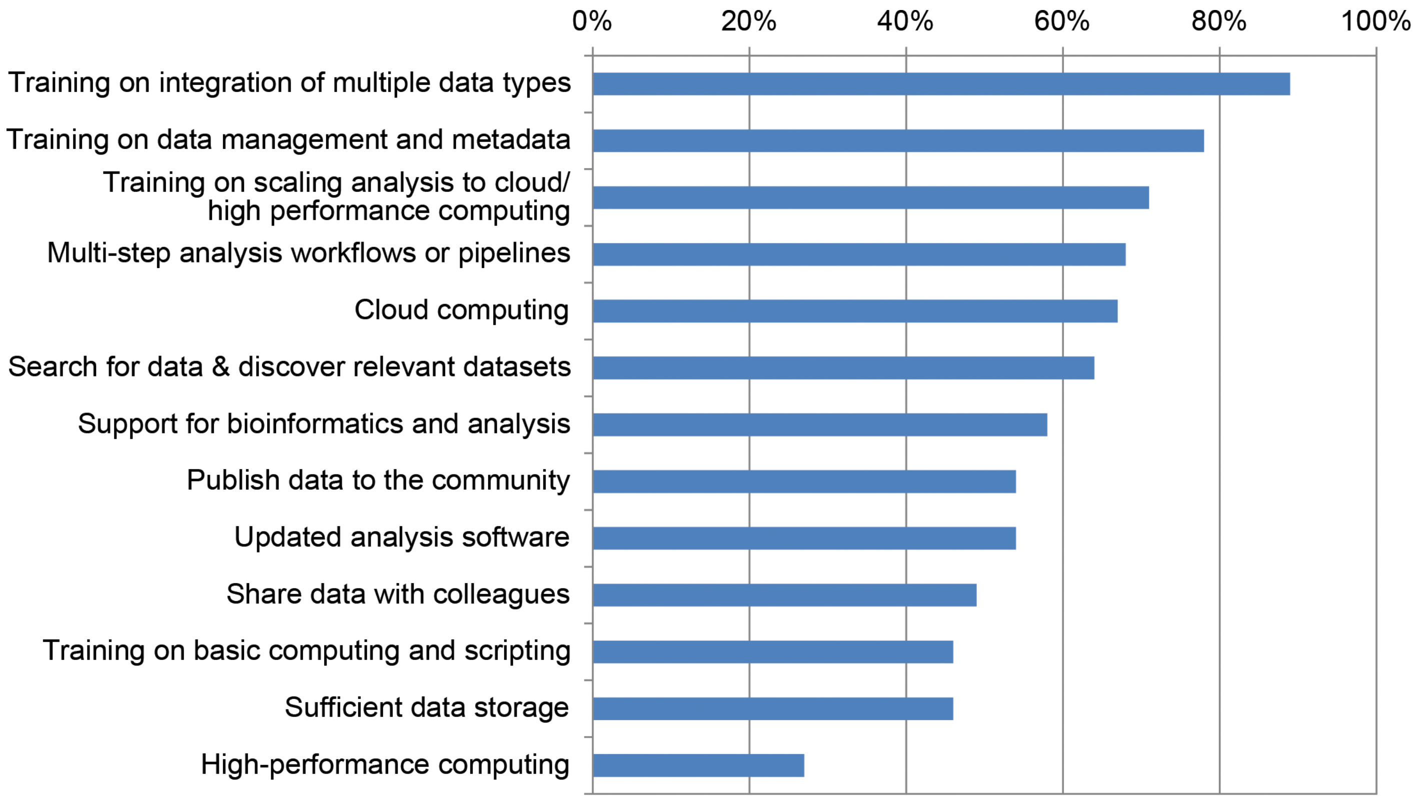

Fig 3. Unmet data analysis needs of National Science Foundation (NSF) Biological Sciences Directorate (BIO) principal investigators (PIs) (percent responding negatively, 318 ≤ *n* ≤ 510).

https://doi.org/10.1371/journal.pcbi.1005755.g003

deeper understanding of "the rules of life" [15, 16]—notably, genotype-environment-phenotype interactions that are essential to predicting how agricultural plants and animals can adapt to changing climates. Such integration demands new standards of data management and attention to metadata about how these data are collected. The BIO PIs in this study are anticipating a new world of pervasive data and the training they will need to become data scientists. Likewise, funding agencies need to recognize that significant new investments in training are now required to make the best use of the biological data infrastructures they have helped establish over the last decade.

Materials and methods

This study was conducted under IRB no. 12–018 from Cold Spring Harbor Laboratory. Working from a list of 5,197 active grant awards, we removed duplicate PIs and those without email addresses to produce a final list of 3,987 subjects. The survey was administered in Survey Monkey using established methods [17]. An initial email invitation with a link to the survey was sent to each subject in June 2016, with 3 follow-up emails sent at 2-week intervals. Surveys were completed by 704 PIs, a response rate of 17.7%, which provided a ±3.35% margin of error at the 95% confidence level.

The respondents were asked to consider 13 computational elements of research, including data storage, discovery, analysis, and sharing. For each need, PIs were asked to reflect on their current use, their anticipated future requirements, and the institutional resources available to meet the need. Data were analyzed in IBM SPSS Statistics version 23. "I don't know" responses were eliminated from the analysis of computational needs questions. Frequencies were calculated for each of the affirmative and negative responses in the computational needs matrix. Chi-square tests for independence were used to determine if there were significant differences in computational needs across the following 3 dimensions: (1) NSF BIO division, (2) research area (bioinformatics/computational biology versus all others), and (3) research group size (groups of less than 5 versus groups with 5 or more).

Data are available for download at https://doi.org/10.1101/108555

Acknowledgments

The authors wish to thank Bob Freeman and Christina Koch of the ACI-REF project for helpful discussions and references during the development of the survey.

References

1. GenBank and WGS Statistics. National Center for Biotechnology Information. 2017. Available from: https://www.ncbi.nlm.nih.gov/genbank/statistics/
2. HiSeq X Series of Sequencing Systems. Illumina. 2017. Available from: https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet-hiseq-x-ten.pdf
3. Wetterstrand K. DNA Sequencing Costs: Data. National Human Genome Research Institute (NHGRI). 2016. Available from: https://www.genome.gov/sequencingcostsdata/
4. Sequence Read Archive. National Center for Biotechnology Information. 2017. Available from: https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=announcement
5. Stephens Z, Lee S, Faghri F, Campbell R, Zhai C, Efron M et al. Big Data: Astronomical or genomical? PLoS Biol. 2015; 13(7): e1002195. https://doi.org/10.1371/journal.pbio.1002195 PMID: 26151137
6. Johnston L. A "Library of Congress" worth of data: It's all in how you define it. 2012 April 25 [cited 8 March 2017]. In: The Signal [Internet]. Available from: http://blogs.loc.gov/thesignal/2012/04/a-library-of-congress-worth-of-data-its-all-in-how-you-define-it/
7. Atkins D. Revolutionizing science and engineering through cyberinfrastructure: Report of the NSF blue-ribbon advisory panel on cyberinfrastructure. 2003. Available from: https://www.nsf.gov/cise/sci/reports/atkins.pdf
8. Duarte AMS, Psomopoulos FE, Blanchet C, Bonvin AMJJ, Corpas M, Franc A et al. Future opportunities and trends for e-infrastructures and life sciences: going beyond the grid to enable life science data analysis. Front Genet. 2015; 6: 197. https://doi.org/10.3389/fgene.2015.00197 https://doi.org/10.3389/fgene.2015.00197 PMID: 26157454
9. Blustain H, Braman S, Katz R, Salaway G. IT engagement in research: A baseline study. Boulder: EDUCAUSE; 2006. Available from: https://net.educause.edu/ir/library/pdf/ers0605/rs/ers0605w.pdf
10. Towns J, Gerstenecker D, Herriott L, Hetrick A, Imker H, Larrison C et al. University of Illinois year of cyberinfrastructure final report [Internet]. 2015. Available from: http://hdl.handle.net/2142/88444
11. Williams J, Teal T. A vision for collaborative training infrastructure for bioinformatics. Ann N Y Acad Sci. 2016; 1387(1): 54–60. https://doi.org/10.1111/nyas.13207 PMID: 27603332
12. Auchincloss L, Laursen S, Branchaw J, Eagan K, Graham M, Hanauer D et al. Assessment of course-based undergraduate research experiences: A meeting report. CBE Life Sci Educ. 2014; 13(1): 29–40. https://doi.org/10.1187/cbe.14-01-0004 PMID: 24591501
13. Rodenbusch S, Hernandez P, Simmons S, Dolan E. Early engagement in course-based research increases graduation rates and completion of science, engineering, and mathematics degrees. CBE Life Sci Educ. 2016; 15(2): ar20–ar20. https://doi.org/10.1187/cbe.16-03-0117 PMID: 27252296
14. Teal T, Cranston K, Lapp H, White E, Wilson G, Ram K et al. Data Carpentry: Workshops to increase data literacy for researchers. Int J Dig Curat. 2015; 10(1).
15. Olds J. Understanding the rules of life: Examining the role of team science. 2015 July 1 [cited 8 March 2017]. In: BioBuzz. Available from: https://oadblog.nsfbio.com/2015/07/01/team_science/
16. Mervis J. NSF director unveils big ideas, with an eye on the next president and Congress. Science. 2016.
17. Dillman D, Smyth J, Christian L. Internet, phone, mail, and mixed-mode surveys. 3rd ed. Hoboken: Wiley; 2009.