A Vision for Collaborative Training Infrastructure for Bioinformatics

Jason J. Williams^a, Tracy K. Teal^b

^a DNA Learning Center, Cold Spring Harbor Laboratory, Cold Spring Harbor NY. ^b Data Carpentry
Address for correspondence: Jason Williams, Cold Spring Harbor Laboratory, 1 Bungtown. Rd. Cold Spring Harbor, NY 11724 USA. williams@cshl.edu

In biology, a “missing link” connecting data generation and data-driven discovery is the training that prepares researchers to effectively manage and analyze data. National and international cyberinfrastructure along with evolving private sector resources place biologists and students within reach of the tools needed for data-intensive biology, but training is still required to make effective use of them. In this concept paper, we review a number of opportunities and challenges that can inform the creation of a national bioinformatics training infrastructure capable of servicing the large number of emerging and existing life scientists. While college curricula are slower to adapt, grassroots, startup-spirited organizations such as Software and Data Carpentry have made impressive inroads in training on the best practices of software use, development, and data analysis. Given the transformative potential of biology and medicine as full-fledged data sciences, more support is needed to organize, amplify, and assess these efforts and their impacts.

Keywords: Bioinformatics; Cyberinfrastructure; Education; Professional Development; Training; Open Science

TECHNOLOGY SCALES, BUT WHAT ABOUT PEOPLE?

As the term “genomical” is suggested as an adjective to replace “astronomical” for the description of biological datasets¹, in 2016 it is self-evident that life science is data science. High-throughput nucleic acid sequencing has given rise to massive amounts of data including repositories such as the *NCBI Short Read Archive (SRA)*. The *SRA*’s 4.5 quadrillion bases of sequence is certain to be an invaluable resource for discovery. Clearly, data production is no longer a bottleneck; genome sequencing costs have decreased 1000-fold in the last decade and this year are undergoing another steep decrease². Advances in image acquisition and analysis promise to accelerate on par with high-throughput sequencing^{3, 4}.

As computation continues to progress, it’s clear that technologies and costs underpinning biological data science will comfortably scale for some time. To quote Sartre however, “*l’existence précède l’essence*” (existence precedes essence) – simply because these data exist we have no guarantee they will generate clear and direct paths to insight.

Francis Collins summarized major lessons from the Human Genome Project 10 years after its completion⁵. His reflections on the remarkable advances in technology are peppered with subtle acknowledgement that some predictions made at the presidential press conference that announced the project’s completion (eight months before actual publication) may have been bold. A decade later, *New York Times* headlines⁶ would capture at least some of the public’s disappointment that promised advances from the sequenced genome are taking longer than originally hoped for.

The *Human Genome Project* was and will continue to be transformative and for Collins, lead to important lessons – including the necessity for open access to data, the need to keep technology a major focus of private-public development, and the importance of wise policy decisions. But is there another lesson to learn, a missing link that could help us avoid another series of heightened expectations for a biology where we can ‘sequence everything’?

While it’s easy to plot trends of how sequencing and computation have progressed, there is a potentially more difficult and more valuable question to ask: *how do we scale the people who use these technologies?* Although this concept paper won’t be able to go much further than outlining current computational resources and framing the question of how we might go about scaling scientists to meet the challenges (and opportunities) of biological data science, we hope to spark discussion and suggest directions meriting further investigation. A vision, if not a hypothesis for the *missing link* that transforms data into insight hinges upon building training infrastructure that scales people. Much of this vision stems the authors’ work to in bioinformatics and to scale training and access to cyberinfrastructure through Software and Data Carpentry and CyVerse.

It’s difficult to speculate on exactly how much progress could have been made if the human genome had landed on a scientific community better prepared for working the prototypical genomics dataset. Still, clues for answering this question will be as relevant to the success (or failure) of biological data science today as they would have been 10 years prior. How we build computational capacity and competence for all levels of life scientists is of everyday relevance. For example, the principal insight of NCBI’s first hackathon on “Advanced Bioinformatic Analysis of Next-Gen Sequencing Data” was that the acquisition and accessibility of these data was “a major hurdle for participants, even though they were genomics professionals.”⁷ If even expert bioinformaticians have difficulties, how will the vast majority of biologists who must use bioinformatics or work with a bioinformatician fare? Additionally, advanced cyberinfrastructure has been developed with both centralized and decentralized solutions, but life scientists are not aware of how to access or utilize these resources. A comprehensive approach to people scaling will involve developing tools and strategies that support large numbers of existing scientists, addressing deficiencies in current training pipelines, and delivering agile approaches for up-to-date and sustainable instruction. We also anticipate the most effective strategies can be informed by the centralized and de-centralized (distributed) approaches that characterize technological infrastructures already driving biological data science.

CENTRALIZING AND DECENTRALIZING –CYBERINFRASTRUCTURE

The development of cyberinfrastructure resources that are both centralized and decentralized, both serves as a model for how training can be established and scaled and provides motivation for this training, as biologists have ready access to the computation needed for their work, but often lack the skills to utilize it effectively. The current development and utilization model for cyberinfrastructure is powerful because by developing and utilizing both centralized and decentralized approaches it simultaneously explores multiple technologies and approaches. We will visit this observation both in thinking about how we use technology and how we train biologists to use technology, outlining available resources and efforts that have made them accessible to researchers.

Biological sensors (sequencers, cameras, etc.) and computational resources have their own blend of centralized and decentralized components; sequencers have moved away from specialized centers and

computation has left the desktop and moved into the cloud. In parallel, cyberinfrastructure has been working in the opposite direction to bring data and compute closer together through public data repositories and resources that bring compute to data. We could condense this model by saying – “find computing where you can, keep instruments close, and your data closer.”

In private-sector computing, virtualization – specifically distributed cloud computing – has made powerful computation accessible and low-cost. One analysis of a genome mapping workflow estimates the costs of mapping a human genome at less than \$50 US, and under \$400 for a scalable production environment on *Amazon Web Services*⁸. Archival data management platforms provide unlimited redundant storage; *Amazon’s* current price is around \$0.007 per gigabyte per month. While exact specifications of private resources like *Amazon Web Services* are not known, they are certainly far more capacious than public-funded initiatives like *CyVerse* (NSF – biological cyberinfrastructure) or even *XSEDE* (NSF – national supercomputing for science)⁹. Though the full potential of these technologies for biology are likely underutilized, they already prompt us to imagine a “computation-unlimited” paradigm for life science.

As much as computation might become unlimited, a completely decentralized approach powered by the private sector contrasts with (but can be complimented by) trends in computing for biological data science. In particular, researchers need access to public data and biological computing has needed to develop cyberinfrastructure suited to searching, sharing, and analyzing biological datasets. Given the need for researchers to work with common data types and formats, centralization (which is some ways more complicated than decentralization) is very efficient. If the tremendously useful and successful *Galaxy* software had a limitation early on, it was that the initial implementation followed a decentralized model in which every user needed to stand up their own *Galaxy* instance. As a result, it was difficult for users to share datasets or avoid upload limitations and it was possible for *Galaxy* admins to mismanage reference data in a way that caused user frustration and/or invalid results¹⁰. Placing most – or at least many of these data resources in one place has been a successful idea since *GenBank* and other biological repositories became web services around the mid-to-late 1990s¹¹. It was not until recently, however, that connectivity and computational capacity could support the next evolutionary step: co-localization of datasets with sufficiently powerful resources for computation and data management.

In 2008, the National Science Foundation funded the *iPlant Collaborative* (now renamed as *CyVerse*) as an answer to biologists’ needs for access to computational tools and sharing datasets – not just at their own institutions, but across the entire community¹². *iPlant* was not just a collection of online tools, but a true cyberinfrastructure with access to cloud and HPC computing as well as mechanisms to manage large biological datasets. Similar infrastructures are being developed including by the European Union’s *ELIXIR* project¹³, *Compute Canada*¹⁴, and the NIH’s *Big-Data to Knowledge* and *Data Commons*¹⁵, all of which set out goals and careful planning to facilitate pairing of data and compute.

Centralized public and decentralized private computing and data are not mutually exclusive; they are complimentary (after a bit of hard work to combine them). Merging public and private resources is very likely the best current solution for meeting the needs of biologists operating at scale who need flexibility to ‘forage’ for the utilization of all available resources. The *Agave API* developed by *CyVerse* and the *University of Texas Austin Texas Advanced Computing Center* combines public and private computing resources in a science-as-a-service platform with the access demanded by power users, and ease-of-use for those who prefer working with graphical user interfaces¹⁶. Both centralized and distributed technologies enable large complex biological investigations to move away from the exclusive operating

domain of international consortia and into the laboratories of individual investigators.

CENTRALIZING AND DECENTRALIZING – TRAINING

Technology and people scale at wildly different rates. Clearly, computing will continue to make it easier for biologists to do big science, but training is needed to make sure that this science will be more powerful and more correct. Troublingly, Gewaltig and Cannon¹⁷ report that the “vast body of results being generated by current computational science practice suffer a large and growing credibility gap: it is impossible to verify most of the computational results shown in conferences and papers.” It is *people*, biologists who lack sufficient training in both the use and best practices of computation who are the bottleneck¹⁸. Are we doing enough to support these biologists? Probably not.

Bioinformatics is a relatively young field, but training efforts and formal curricula have been around for more than a decade^{19,20}. Still, without a large distributed body of faculty comfortable with bioinformatics on the same level as other (more biology-based) techniques like molecular cloning, a large skills gap will continue to exist. Survey data clearly indicate that the majority of biologists don't possess the skills needed to marshal and manage computing resources needed for their investigations²¹. Survey data from *EMBL* state that 60% of biologists report a need for more training (compared to 5% who say they need more computing power)²². Similar *CyVerse* survey data convey that greater than 95% of researchers are or plan to be working with large genomics datasets, but 68% state they have only beginner bioinformatics skills or less²³.

Appropriately, training efforts examined as centralized and decentralized approaches may highlight how we might borrow and adapt from the successful technological model – could we “keep training local, but act to develop training materials globally”? Some researchers are able to take advantage of centralized training experiences at their own institutions. A local group of faculty and students, often coalescing around a new faculty member(s) who will ‘introduce genomic approaches’ can be an excellent way to help bring a department up to speed. A disadvantage to this approach is that it while it may be very good, it can also in some cases be unhelpful – especially if the training program is isolated from the community and propagates out-of-date information. Local training efforts make an impact, but sometimes they don't even manage to scale beyond a department (great training may be happening in microbiology, but is totally unobtainable in ecology). Additionally, centralized approaches are almost necessarily slower to react to new technologies which again scale orders of magnitudes faster than people or curricula.

Massive Open Online Courses (MOOCs) and other online teaching and training are an existing approach that can definitely scale the number of learners. However, since course development is typically not open, they still suffer from the same of the same problems as any training that originates from a single source (one university or one small group of instructors). Additionally, we know that MOOCs have very low (<10%) completion rates²⁴ – though these rates may be better for professional development learning.

Collaborative, decentralized training has clear advantages for democratizing access and responding at scale to rapidly evolving science. *Software Carpentry* and *Data Carpentry* have developed reputations for using these approaches to deliver training and present models that more closely resemble Silicon Valley software startups than university departments. *Software Carpentry* started in 1998, then delivering five-day workshops on software development for science. Their focus has been to develop

skills and strategies for making science more reproducible and efficient by giving scientists skills to develop automated workflows, use version control, and test the software they produce. In 2014 *Data Carpentry* took a similar approach, but with a focus on data analysis, management, and domain-specific topics in biology – the reader is strongly encouraged to consider Teal et al.²⁵ for a discussion on needs assessment for biologists, and a deeper look into the thinking behind these approaches. In a sense, the *Software* and *Data Carpentry* models borrow the best aspects of formal and peer-training by scaling it out – their instructors are all volunteers, who have expertise and work within their respective scientific domains. Importantly, instructors share knowledge, experience, and pedagogical technique because they are unified by an open and collaborative lesson development infrastructure. Lessons are effectively developed via a crowdsourcing approach involving moderation by community selected lesson maintainers, and contributions are organized through GitHub (again borrowing from the ideas of modern software development). The result is that lessons are openly reviewed, iteratively developed, and capture the consensus of community knowledge. So far, with a pool of just under 500 certified instructors, operating internationally at workshops on a weekly basis for the last several years, *Software Carpentry* has been able to reach more than 17,000 learners.

Many other decentralized (in this case decentralized meaning they are driven by external and internal instructors following a dynamic and collaboratively developed curriculum, rather than an isolated group of practitioners) bioinformatics training programs exist (e.g. Cold Spring Harbor Courses, Marine Biological Laboratory, STAMPS, NGS, EDAMAME, etc.), and it would be a worthy effort to align these decentralized efforts into scalable programs that share resources and increase accessibility. Even as they continue to scale, *Software* and *Data Carpentry* are only the beginning of a possible solution. There continues to be vast unmet demand for their services (typically workshops fill up within hours of opening registration). Furthermore, while two-day or even week-long workshop models can communicate important basic skills to get biologists started with effective bioinformatics, they are only a good beginning - accommodating all the necessary skills or advanced training needs will require sustained longer-term efforts.

DEVELOPING A TRAINING INFRASTRUCTURE FOR BIOINFORMATICS

With sustained advances in data acquisition and analysis and successful ideas for training, what are some ideas and visions for what can be done next? We present a few suggestions driven by the authors' experience in training in bioinformatics –in the context of delivering professional development at scale.

1) *Technological infrastructure must be complemented by training infrastructure.* There is a tremendous opportunity to create a setting where the tools of bioinformatics can be as standardized as other scientific instrumentation. As capacity and technologies scale, cyberinfrastructure and cloud resources make it possible not only to analyze big data, but will increasingly allow biologists to work from common platforms and interfaces. For example, technology like *Docker Containers*²⁶ can help us deliver training environments that will make lesson materials applicable across diverse contexts. It will be critical for cyberinfrastructure platforms to be closely tied to training and user needs. The need for biologists-friendly designs discussed in Kumar and Dudley¹⁸ are very relevant, and one of the strategies CyVerse has employed in making its platforms more usable is to pair conventional documentation with *Software* and *Data Carpentry* training on best computational practice. Amplifying the *EMBL* survey data, it is necessary to here again state that training must be a priority and not an afterthought. It's probably asking too much to suggest that a bright graduate student should be concerned with how to train thousands of potential users before she publishes a novel algorithm. Collectively, however, we can

develop training infrastructure that makes it easy to onboard new technologies using modular approaches, templates, and tools for training materials that can be delivered and powered by community effort.

2) We need to brainstorm on methods to assess both the productivity and efficiency of biologists and our training efforts. This is perhaps one of the more difficult goals which the authors are aware of. Even at the earliest level, it is probably true that we are doing a poor job at assessing how we first teach bioinformatics. A meta-analysis of the assessment efforts of 226 papers on genomics and bioinformatics education reveals that less than 10% of these studies contained any evidence of the reliability or validity of the assessment approaches used²⁷. There are tremendous bodies of work and expertise on evidence-based instruction, but we need to better understand how we could leverage these to develop the best training. Without delving into the largely open question of how to measure productivity in science, we could probably develop better tools to determine if scientists are adopting practices that should ensure greater productivity and reproducibility. If a graduate student is working with a complicated whole-genome association study using an *Excel* spreadsheet rather than an appropriate *R* data frame, they are probably being hamstrung in very significant ways. Explicit attention to this goal should be able to identify and leverage current relevant efforts. A reasonable outcome could include improved guidelines for data management plans that require more explicit details for reporting on bioinformatics methods in a way that reflects minimum standards for reproducibility.

3) We should advocate for Bioinformatics Core Competencies. There have been broad efforts to develop core competencies for bioinformaticians and bioinformatics training efforts²⁸ and examples such as the newly-funded *Network for Integrating Bioinformatics into Life Sciences Education* (NIBLSE)²⁹ indicate growing awareness of this need. With community support, a highly communicated effort to develop general recommendations and core competencies would establish clear touchstones for biologists who want to know if they can trust their skills and interpretation of results (or identify where further expertise is advisable). Aligning training to core competencies also allows biologists to focus learning on topics most relevant to their research. On a policy level, this type of effort was employed to improve undergraduate biology education through the *NSF and AAAS's Vision and Change* report³⁰. A “living” set of community-developed bioinformatics lessons that represent core competencies in several relevant subdomains (e.g. genome assembly, GWAS, image acquisition, etc.) would be a powerful resource for delivering be up-to-date training at scale.

4) We need to think about scaling the peer-training model. Besides collaborative lesson development, *Software and Data Carpentry* are examples of how volunteer trainers (who are practicing professionals in their own domains) can be important resource for improving how skills are disseminated across communities. These instructors are passionate about sharing their expertise, and are given further *Carpentry* training on how to be effective teachers. Although the biology community has these human resources, “the scientific community has failed to craft attractive career paths for those who do the analyses it increasingly requires.”³¹ We need to find a way to harness community enthusiasm and ingenuity as part of the dissemination process. This may take a variety of forms – from exploring ways to hybridize in-person workshops with virtual training to creating generic training infrastructure so that high-value training can be regularly distributed across institutions. All state universities for example, could host bioinformatics training nodes that would function as a “content delivery network” for bioinformatics. Additionally, we also need to understand how to incentivize and create career paths for these human resources.

5) We need to encourage local ‘communities of practice.’ While training is essential, it doesn’t stop once a researcher leaves the room. Training is an ongoing process that can be furthered and encouraged by engagement with a community working in and developing these practices. Programs like *Python* or *R* user groups and meetups are successful in bringing people together with the same interests, welcoming new users and presenting information on new approaches or responding to questions from the community. The local aspect of these groups means that researchers have local resources for help and encouragement. Additionally, they develop peer-incentive for good practice in computation. Encouraging the development of these groups, by incentivizing their leaders, providing easy-to-use resources like ‘data analysis book clubs’ or information on best-practices for community organization could further these efforts.

CONCLUDING REMARKS

Not every biology problem will require massive parallel computation and access to terabytes of data, but we need to work towards ensuring that when these resources can help, they are being exploited. Improving the educational pipeline to create greater numbers of formally trained bioinformaticians will be important, but that approach probably misses the importance of creating a community of biologists who are computation-informed. As Lincoln Stein notes, “Bioinformatics has become too central to biology to be left to specialist bioinformaticians.”³² It is the opinion of the authors that these approaches hold potential to scale biologists. The vision presented here along with many more ideas from bioinformaticians and educators in the field will require significant support to be crafted into a true training infrastructure, perhaps at a level of effort complementary to the resources invested in developing computing infrastructure. Biologists don’t need to become computational experts, but must have shared vocabulary and skills needed to be conversant with experts. A national bioinformatics training infrastructure may be the best strategy to empower researchers to participate in biology’s evolution as a data science. Approaches that leave people behind will mean that the coming yottabytes of biological data could amount to adding hay to the proverbial haystack; we risk burying insights.

ACKNOWLEDGMENTS

The authors thank Lindsay Barone, Jonah Duckles and Greg Wilson for helpful comments and corrections.

DISCLOSURE STATEMENT

The authors both have relationships with CyVerse and Data Carpentry (Jason W., and Tracy T. respectively). While we have no-financial benefit or interest conflict to declare, we want to make it obvious to the reader that our perspective’s on these projects is an internal one.

CITATION FOR ACCEPTED PRINT VERSION:

Williams, Jason J. and Teal, Tracy K. (2016), A Vision for Collaborative Training Infrastructure for Bioinformatics. *Annals of the New York Academy of Sciences*. doi: 10.1111/nyas.13207

REFERENCES

1. Stephens, Z.D., S.Y. Lee, F. Faghri, et al. 2015. Big Data: Astronomical or Genomical? *PLOS Biology*. 13: e1002195.
2. K.A. Wetterstrand. 2015. DNA Sequencing Costs- *Genome.gov*, 2015. Accessed November 25, 2015. <http://www.genome.gov/sequencingcosts>.
3. Fahlgren, N., M.A. Gehan & I. Baxter. 2015. Lights, camera, action: high-throughput plant phenotyping is ready for a close-up. *Current Opinion in Plant Biology*. 24: 93-99.
4. Chapnick, D.A., J. Jacobsen & X. Liu. 2013. The Development of a Novel High Throughput Computational Tool for Studying Individual and Collective Cellular Migration. *PLoS ONE*. 8: e82444.
5. Collins, F. 2010. Has the revolution arrived? *Nature*. 464: 674-675.
6. N. Wade. 2015. A Decade Later, Human Genome Project Yields Few New Cures - *NYTimes.com*. June 13, 2010. Accessed November 25, 2015. <http://www.nytimes.com/2010/06/13/health/research/13genome.html?pagewanted=all>.
7. NCBI Staff. 2015. NCBI's First Hackathon: Advanced Bioinformatic Analysis of Next-Gen Sequencing Data, *NCBI Insights*. May 22, 2015. Accessed November 25, 2015. <http://ncbiinsights.ncbi.nlm.nih.gov/2015/05/22/ncbis-first-hackathon-advanced-bioinformatic-analysis-of-next-gen-sequencing-data/#more-613>.
8. Fusaro, V.A., P. Patil, E. Gafni, et al. 2011. Biomedical Cloud Computing With Amazon Web Services. *PLoS Comput Biol*. 7: e1002147.
9. Juve, G., M. Rynge, E. Deelman, et al. 2013. Comparing FutureGrid, Amazon EC2, and Open Science Grid for Scientific Workflows. *Comput. Sci. Eng.* 15: 20-29.
10. Blankenberg, D., J.E. Johnson, J. Taylor, et al. 2014. Wrangling Galaxy's reference data. *Bioinformatics*. 30: 1917-1919.
11. Smith, K. 2013. A Brief History of NCBI's Formation and Growth: *National Center for Biotechnology Information (US)*.
12. Goff, S.A., M. Vaughn, S. McKay, et al. 2011. The iPlant Collaborative: Cyberinfrastructure for Plant Biology. *Frontiers in Plant Science*. 2.
13. Publication Office of the European Union. 2011. Strategy Report on Research Infrastructures - Roadmap 2010. 2011. Accessed November 25, 2015. https://www.elixir-europe.org/system/files/esfri-strategy_report_and_roadmap.pdf.
14. Canada Foundation for Innovation Cyberinfrastructure Initiative. 2014. Draft Calls for Proposals For Consultation. Accessed November 25, 2015. <http://www.innovation.ca/sites/default/files/Funds/cyber/Cyberinfrastructure%20Initiative%20-%20Draft%20call%20for%20proposals.pdf>.
15. Bourne, P. 2014. The Commons. October 7, 2010. Accessed November 25, 2015. <https://pebourne.wordpress.com/2014/10/07/the-commons/>.
16. Dooley, R., Vaughn, M., Terry, S., Stanzione, D., & Skidmore, E. 2012. Software-as-a-Service: The iPlant Foundation API [abstract] *IEEE Workshop on Many-Task Computing on Grids and Supercomputers (MTAGS)*.
17. Gewaltig, M.-O. & R. Cannon. 2014. Current Practice in Software Development for Computational Neuroscience and How to Improve It. *PLoS Comput Biol*. 10: e1003376.

18. Kumar, S. & J. Dudley. 2007. Bioinformatics software for biologists in the genomics era. *Bioinformatics*. 23: 1713-1717.
19. Ranganathan, S. 2005. Bioinformatics Education—Perspectives and Challenges. *PLoS Comp Biol*. 1: e52.
20. Tan, T., S. Lim, A.M. Khan, et al. 2009. A proposed minimum skill set for university graduates to meet the informatics needs and challenges of the "-omics" era. *BMC Genomics*. 10: S36.
21. Dudley, J.T. & A.J. Butte. 2009. A Quick Guide for Developing Effective Bioinformatics Programming Skills. *PLoS Comput Biol*. 5: e1000589.
22. EMBL. 2013. BRAEMBL Community Survey Report - 2013 | EMBL Australia Bioinformatics Resource. Accessed November 25, 2015. <http://braembl.org.au/news/braembl-community-survey-report-2013>.
23. Williams, J.J. 2015. Unpublished Survey Data – iPlant Collaborative
24. Gütl, C., R.H. Rizzardini, V. Chang, et al. 2014. "Attrition in MOOC: Lessons Learned from Drop-Out Students". In *Communications in Computer and Information Science*: 37-48. Springer Science + Business Media.
25. Teal, T.K., K.A. Cranston, H. Lapp, et al. 2015. Data Carpentry: Workshops to Increase Data Literacy for Researchers. *International Journal of Digital Curation*. 10.
26. Boettiger, C. 2015. An introduction to Docker for reproducible research. *ACM SIGOPS Operating Systems Review*. 49: 71-79.
27. Campbell, C.E. & R.H. Nehm. 2013. A Critical Analysis of Assessment Quality in Genomics and Bioinformatics Education Research. *Cell Biology Education*. 12: 530-541.
28. Welch, L., F. Lewitter, R. Schwartz, et al. 2014. Bioinformatics Curriculum Guidelines: Toward a Definition of Core Competencies. *PLoS Comput Biol*. 10: e1003496.
29. NSF. 2015. NSF Award Search: Award#1539900 - RCN-UBE: Network for Integrating Bioinformatics into Life Sciences Education (NIBLSE). Accessed November 25, 2015. http://www.nsf.gov/awardsearch/showAward?AWD_ID=1539900&HistoricalAwards=false.
30. AAAS. 2009. Vision and Change in Undergraduate Biology Education. Accessed November 25, 2015. <http://visionandchange.org/>.
31. Chang, J. 2015. Core services: Reward bioinformaticians. *Nature*. 520: 151-152.
32. Stein, L.D. 2008. Bioinformatics: alive and kicking. *Genome Biol*. 9: 114.