A voice for bioinformatics

Carrie L. Ganote*
National Center for Genome Analysis Support
2709 E. 10th St.
Bloomington, Indiana 47408
cganote@iu.edu

Sheri A. Sanders*
National Center for Genome Analysis Support
2709 E. 10th St.
Bloomington, Indiana 47408
ss93@iu.edu

Bhavya Nalagampalli Papudeshi*
National Center for Genome Analysis Support
2709 E. 10th St.
Bloomington, Indiana 47408
bhnala@iu.edu

Phillip D. Blood†
National Center for Genome Analysis Support
300 S Craig St.
Pittsburgh, Pennsylvania 15213
blood@psc.edu

Thomas G. Doak*
National Center for Genome Analysis Support
2709 E. 10th St.
Bloomington, Indiana 47408
tdoak@iu.edu

ABSTRACT

One of the challenges to adoption of HPC is the disjunction between those who need it and those who know it. Biology (specifically, genomics) is a growing field for computational use, but the typical biologist does not have an established informatics background. The National Center for Genome Analysis Support (NCGAS) aids users in getting past the initial shock of the command line and guides them toward savvy cluster use. NCGAS is initiating a push to become domain champions alongside Oklahoma State's Brian Cougar. Our position at IU gives us a close relationship with XSEDE and we already fulfill a role in pushing users toward XSEDE resources when our local clusters are ill-suited to the job. We currently act as liaison between biologists and Jetstream, IU and TACC's research computing cloud. Typical issues include: Software installation; Software usage - what parameters do I choose, and how do I interpret the results; Batch job submission; Understanding how queues and job handlers work; Data movement, Spinning up VMs on Jetstream We will discuss how we have structured our support, and illustrate our impact on XSEDE resources.

CCS CONCEPTS

• **Applied computing** → *Genomics*;

KEYWORDS

Genomics, Bioinformatics, User support

ACM Reference format:
Carrie L. Ganote, Sheri A. Sanders[1], Bhavya Nalagampalli Papudeshi[1], Phillip D. Blood, and Thomas G. Doak[1]. 2017. A voice for bioinformatics.

*Indiana University
†Pittsburgh Supercomputing Center, Carnegie Mellon University

PEARC17, July 2017, New Orleans, Louisiana USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

In *Proceedings of Practice & Experience in Advanced Research Computing 2017 Conference, New Orleans, Louisiana USA, July 2017 (PEARC17),* 5 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION TO NCGAS

NCGAS[2][4] is a collaborative project between the lead institution, Indiana University (IU), and the Pittsburgh Supercomputing Center (PSC) at Carnegie Mellon University. Following an ABI Development award that established NCGAS, we are now funded under an ABI sustaining grant (NSF Awards DBI-1458641 and ABI-1062432:).

At IU, NCGAS is part of the Indiana University Pervasive Technology Institute (PTI) and has significant HPC facilities, human resources, and administrative support from PTI and IU. Likewise, NCGAS-funded collaborator PSC maintains extensive HPC resources and supporting services. During the second year of this sustaining award, NCGAS has continued to make significant strides in using NSF fundingâĂŤwith additional funding and facilities from IU and PSCâĂŤto aid discovery and innovation in the biological sciences in the US. Under the direction of IU, NCGAS has developed new opportunities through collaborative efforts between IU and PSC and has continued to aid in discoveries that range from a better understanding of basic biological processes, to discoveries that will aid management of economically and ecologically important animals and plants.

NCGAS continues to assist NSF researchers in genomics research. From the beginning we have accomplished this by forming a "supply line" from the researcher's specific data and questions to HPC hardware, specialized applications and knowledge. Some researchers only need access to large memory clusters, which we can provide in a number of ways; others need instruction in basic HPC use and genomic analysis. We have been successful in this and continue to attract new users, often by word-of-mouth (documented in our just closed survey of users). One of our on-going tasks is to stay current: new hardware becomes available, state-of-the-art applications change, new data types become available (we are starting to see a significant number of PacBio data sets), and researchers change. For example, this year we installed or updated five packages (spades, hisat2, salmon, STAR, kallisto) for assembly and analysis of

PacBio long-read sequencing data. Hybrid assemblies are quickly becoming popular, and we installed Canu, MaSuRCA and PacBio's SMRT analysis software. Funded collaborator PSC also makes many of these packages available on PSC systems, and also focuses on enabling high-quality metagenome assembly and analysis.

2 WHY ARE BIOLOGISTS SPECIAL?

High Performance Computing (HPC) has been a game-changer for many fields in mathematics and the sciences. In a traditional lab setting, desktop computers are used for the analysis of large data; this can take weeks to run, is prone to crashing, and ties up the entire machine, which is often shared among lab members. HPC fast-forwards this process immensely and can take the computational bottleneck away, allowing more agile experimentation by the researcher. The driving fields for HPC, however, have traditionally been physics and math - as can be witnessed by the ubiquitous measure of HPC in FLOPS (floating point operations per second). Biologists are relative newcomers to HPC and often have different needs than other domains.

Genomics in particular is a field now awash in the âĂIJbig data revolutionâĂİ. Sequencing chemistry is changing so fast that algorithms to handle it can hardly keep pace. Development of new technologies for sequencing genomes, incorporating an increasing variety of other omics data into the existing frameworks, and the additional burden of a lack of mature standards for file formats and file handling coalesces into a famine of proper informatic support for the science. In addition, data is more heterogeneous than other "big data" domains (for example astronomy) and can require long term storage of raw data, high memory, or numerous CPUs[3].

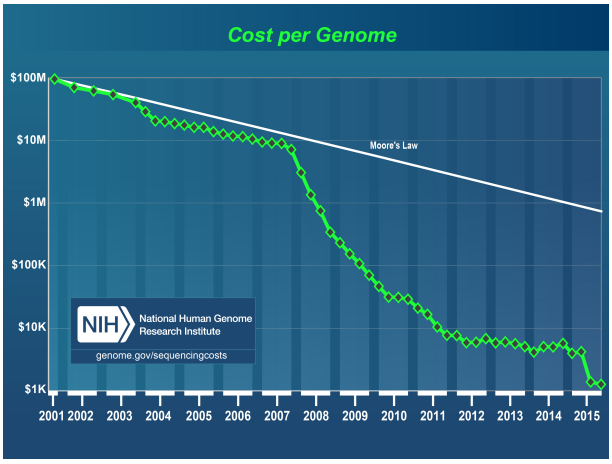

Figure 1: The cost of data acquisition in genomics is falling much faster than Moore's Law. As a result, scale of experiments is increasing and storage and computational resource needs are increasing as well. An estimated 1 zettabase/year of sequence data (not including downstream analysis) will be produced by 2025 [3]. Image from NIH, 2016[1].

Software development to handle the rapid changes in the technology from the lab is often done in the lab. Biologists, both students and PIs, traditionally trained on bench work, find themselves needing to pick up informatics "on the side," in order to make headway in their research. Often, software choice for analysis is influenced heavily on the ease of installation or use, rather than the biological rigor or computational efficiency. On the other side of the equation, computer scientists supporting software development for the biology domain can have issues producing something that will actually be biologically relevant to the researcher.

Dealing with the explosion in data, the fast pace of technology change, high resource needs, and inexperienced end users in biological analysis requires national level, domain specific bridges to computational expertise. NCGAS strives to fill this need by providing computational resources (storage and computation) to NSF funded genomics projects, providing expertise in current technology and software, and training in basic computation (UNIX, software, clusters, etc) - all listed by users as needed in their research programs (Figure 2).

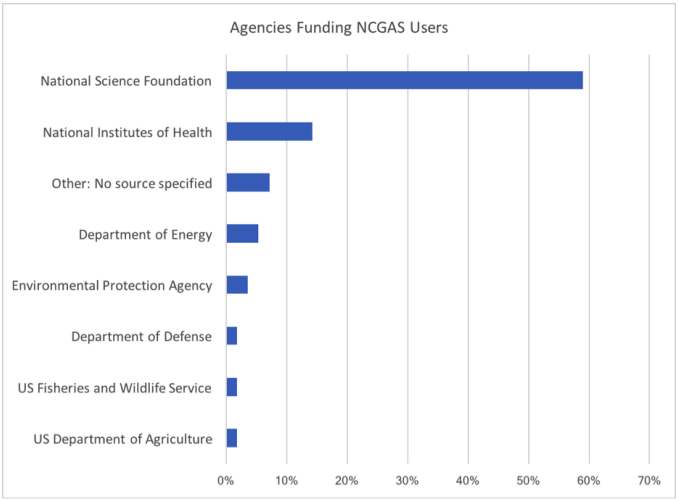

Figure 2: Response by NCGAS users to 2017 annual survey question âĂIJwhich of these would be helpful to you?âĂİ

3 OUR ROLE IN XSEDE AS A DOMAIN CHAMPION (AND LEVEL 3 SERVICE PROVIDER)

If informatics is embraced as a core concept in undergraduate or even high school biology curriculum, this could mean a rapid increase in the number of requests for help by new students as well faculty suddenly finding themselves needing to teach things they themselves do not know. As a fairly small Center, we don't have the resources to conduct this training at IU, let alone the nation. Our greatest impact comes when there are specific groups and projects that require intense one-on-one help for relatively short periods of time. One way that we boost our effectiveness is to target our peers in Advanced CyberInfrastructure and lend them some of our expertise, when they might have a problem that is too specific to genomics/bioinformatics for the facilitator to be able to easily

answer. In this way, we train trainers and scale up our efforts without severely limiting our ability to maintain our primary user-care focus.

The Campus Champion community has made heroic efforts to bring HPC closer to the user. Indiana University has two campus champions within Research Technologies, and our niche is not so broad that we would be useful when helping other disciplines. NCGAS is extremely domain-focused; while we believe that HPC should be useful to research and to the public it serves, the technology is secondary to the needs of biologists. If biology as a field were to outgrow the need for traditional high performance computing, our facilitation would move with it.

The domain champion is a great fit for NCGAS as it allows us to reach a broad group of people and plays well to our strengths. In addition, we learn from the experience of other centers through their problems and their users' needs. As a group, we try our best to distill the vast amount of information and options down to a sensible best practices for the field. Increased cross-communication between HPC centers, domain experts, software developers, and biologists new to informatics will make possible the most optimal fit of user needs to hardware and software available.

NCGAS has just started our campaign as a domain champion, so we haven't had much time/change yet to fully realize its impact on us or our impact on the community. We intend to become more involved with the existing national community to enable more symbiosis between the existing nsf funded XSEDE project and our own nsf funding. Efficient use of government funds will be beneficial to the scientific community in years to come, to prove our continued positive impact to public interests.

4 OUR ROLE AS JETSTREAM LIAISONS TO BIOLOGISTS

The NSF-funded cloud environment JetstreamâĂŤwhile not providing very large memoryâĂŤhas already helped NCGAS researchers accomplish genomics science (ex. ecological-genomics projects from AR, done by the labs of Michael and Marlis Douglas, and aided by UofA Jeff Pummill, AR High Performance Computing Center), and we will continue to expand its uses.

Jetstream is playing an increasingly important role in our collaboration with PI Keithanne Mockaitis on several plant transcriptomics projects. We have brought up a persistent web host using Jetstream's API in order to keep a reliable asynchronous communication platform available for the various projects in the form of a wiki. The same instance hosts JBrowse instances for the Coffee transcriptome project, using XSEDE's Wrangler system to provide storage for genomes on the back end.

In conjunction with the Jetstream development team, we have just completed a survey of field and marine station directors and managers, as represented by the membership of the Organization of Biological Field Stations (OBFS; http://www.obfs.org/), an association of more than 200 field stations and professionals concerned with field facilities for biological research and education. The survey's intent was to ascertain stations' general cyberinfrastructure needs, and specifically how the Jetstream cloud could serve their needs.

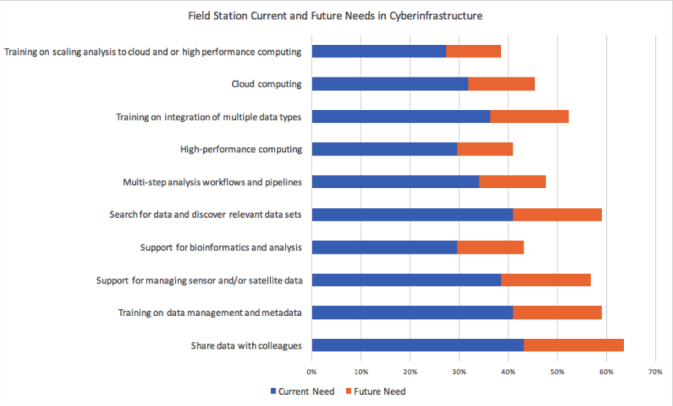

Figure 3: Cyberinfrastucture needs of field and marine stations, as per XSEDE survey 2017

5 EXAMPLES OF NCGAS SERVICES

NCGAS has participated in outreach and education opportunities since its inception. Our methods include REU internships, volunteering at local non-profit education fairs, supporting courses through infrastructure and guest lectures, providing training and compute power for multi-day workshops hosted at Indiana University and elsewhere, giving conference talks and posters, and collaborating with international research groups.

5.1 *Supporting Workshops*

NCGAS (National Center for Genome Analysis and Support) supports extended workshops in genomic analysis, such as the Environmental Genomics Workshop at Mount Desert Island Biological Laboratories in Bar Harbor, Maine. This is an extensive workshop providing training in cluster use, bioinformatic analysis, and data management.

In 2016, nine students and nine postdoctoral fellows and faculty from 14 universities learned to design experiments, create 96 RNA-seq libraries from water fleas (Daphnia pulex), and analyze the data using IU's Karst compute cluster. The data produced during this analysis is used in ongoing research on evolution and heavy metal response and serves as preliminary data for several investigations.

As research in genomics becomes increasingly dependent on large sequencing data, more and more biologists are required to learn how to use high performance computing and bioinformatic programs. These analyses and software are relatively new, but very powerful tools. Most biology graduate students lack the necessary computational training, and many professors and post-docs are new to these methods. Training programs such as this are integral to supporting a broad scope of biological questions. Participant attended the workshop seeking to apply the training to projects ranging from conservation of coastal dolphins to investigation into the causes of miscarriage in humans.

5.2 *Guest Lectures and Course Support*

NCGAS works with universities that don't have large scale computing facilities to support classes and training of biologists. For

example, NCGAS has been working with Bethune-Cookman University, an HBCU (Historically Black Colleges and Universities) in Daytona Beach, Florida to support the new bioinformatics course titled âĂIJAdvanced Computing Resources in BiologyâĂİ. Led by Dr. Raphael D. Isokpehi, students learned the principles of command line genomics software and python programming as well as gained exposure to large-scale computing, using IU's Mason Compute Cluster as learning space.

The students also attended a guest lecture by staff from the National Center for Genomic Analysis Support (NCGAS) at Indiana University. Four students met with NCGAS at the 2016 Supercomputing Conference in Salt Lake City, UT. They were able to inquire about the bioinformatics career path from industry professionals and gain insight into what to expect when moving from biology heavy training to computational-based science.

Bioinformatics is a fast growing field, but training requires university level training in cluster computing and genomics software. âĂIJThe collaboration with NCGAS is enabling biology students at Bethune-Cookman University to access computing resources for working effectively with large-scale biological dataâĂİ says Dr. Isokpehi. Students are also acquiring the expertise to compete for internships and fellowships in pursuit of a career in biology or bioinformatics. Providing science and engineering cyberinfrastructure and training to Bethune-Cookman University and other institutions helps expand the curriculum available without the overhead cost of institution-level scientific computing infrastructure. In person interaction with students and new users at conferences solidifies relationships and increases NCGAS's presence in the community. As the course was successful in providing previously inaccessible resources for training students, NCGAS will continue working with Bethune-Cookman in this capacity this Spring 2017 semester.

5.3 *Collaborations and Research Support*

NCGAS provides analytical and support services to collaborators that include staff on grants. One such example is a long term collaboration with Cenicafe , based in Columbia and Cornell. The mission of Cenicafe is to improve the livelihoods and outcomes for coffee producers in Columbia, with particular focus on combatting coffee rust, a disease which threatens coffee growers and may percolate through the industry, drying up supplies for coffee consumers. NCGAS works with Cenicafe to analyze genomic information and host the results on genome browsers for easy interpretation. This collaboration is partially funded by grants to support investigating transcriptomics of the Coffea arabica plant. Meet-ups are planned around meetings, such as Plant and Animal Genomics, allowing for discussions with the collaborators to solidify goals for the next year, clarify the progress from the last year to all stakeholders, and strategize about how to proceed.

5.4 *Software Support and Collaboration*

The Broad Institute's messenger RNA (mRNA) assembler, Trinity (developed by Aviv Regev, Brian Haas and others), is one of the most popular tools for assembling short reads (100-150 base pairs) of mRNA into the transcripts (200-15000bp) produced by an organism. According to recent Github download stats, over 16,000 people have downloaded the program 4.4 million times since 2015, and this is increasing every year, with 1000 downloads to unique IP addresses per month in 2017, thus far, up from 770/month in 2016 and 500/month in 2015. This trend is interesting, especially in light of the heavy use of shared installations of Trinity on HPC resources, which only count as one download, despite being used by potentially hundreds of individuals. The Trinity software requires 1GB of

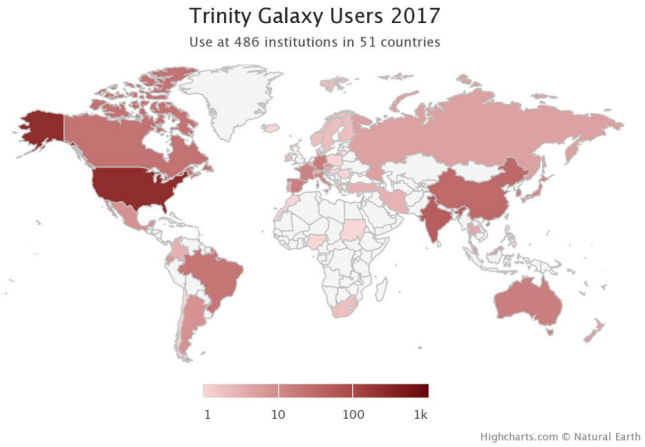

Figure 4: Map of self reported user locations for Trinity Galaxy as of May 2017. Scale is logarithmic based on density. Map created using highcharts.

RAM for every million reads of sequence, which means a single lane of Illumina sequence (a low end for many projects) requires 180GB of RAM. With sequencing costs continuing to decrease, the scale of mRNA sequencing projects has increased - resulting in many lanes of Illumina and TB-scale memory requirements. Therefore, Trinity requires HPC compute resources, such as IU's Mason or through web interfaces such as the Trinity CTAT Galaxy, maintained by the National Center for Genome Analysis Support (NCGAS) at IU. Trinity CTAT Galaxy alone is being used by 665 users across 486 institutions in 51 countries (see map), averaging about 130 jobs per monthâĂŤagain, this high-volume use only accounts for a couple of the 4.4 million downloads of the software.

When hundreds of people are using a single instance of a software for very diverse projects, issues and complications become evident much faster than when single users are using individual local installations. NCGAS provides user support, and agglomerates user issues into direct feedback to the Broad Trinity developers. This partnership between developers and user-facing centers like NCGAS contribute significantly to the continued success of software, as it becomes more efficient and better handles biological complexities.

Earlier in NCGAS's partnership with Trinity, NCGAS made improvements boost speeds by a factor of 4. Now NCGAS is working to seamlessly pass jobs to different clusters to handle TB scale memory jobs in a timely manner.

6 SUMMARY AND FUTURE EFFORTS

NCGAS was founded with the goal of bringing HPC and genomic services to biologists, and we feel we have succeeded in this effort,

although there is much more to be done. But we also serve to bring biologists to national HPC resources, especially the many systems funded by NSF: while biologists in the the age of genomics need HPC like never before, they have only slowly discovered the many free resources available to them. NCGAS serves are a domain-specific entry to these resources and serves as an XSEDE partner in this effort.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Yuhua Li for providing the matlab code of the *BEPS* method.

The authors would also like to thank the anonymous referees for their valuable comments and helpful suggestions. The work is supported by the National Natural Science Foundation of China under Grant No.: 61273304 and Young Scientsts' Support Program (http://www.nnsf.cn/youngscientsts).

REFERENCES

[1] 2016. *The Cost of Sequencing a Human Genome - National Human Genome Research Institute.* https://www.genome.gov/sequencingcosts/.

[2] Richard D. LeDuc, Le-Shin Wu, Carrie L. Ganote, Thomas Doak, Philip D. Blood, and Matthew Vaughn. 2013. National Center for Genome Analysis Support Leverages XSEDE to Support Life Science Research. In *Proceedings of the Conference on Extreme Science and Engineering Discovery Environment: Gateway to Discovery.* ACM Press.

[3] Zachary D. Stephens, Skylar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Ravishankar Iyer, Michael C. Schatz, Saurabh Sinha, and Gene E. Robinson. 2015. Big Data: Astronomical or Genomical? *PLOS Biology* 13, 7 (July 2015), E1002195.

[4] Craig A. Stewart, William K Barnett, Matthew W. Hahn, and Michael R. Lynch. 2015. ABI Development: National Center for Genome Analysis Support. *PTI Technical Report PTI-TR15-009* (December 2015).