National Center for Genome Analysis Support Leverages XSEDE to Support Life Science Research

Richard D. LeDuc
Indiana University
2709 E. Tenth Street
Bloomington, IN 47408
rleduc@iu.edu

Le-Shin Wu
Indiana University
2709 E. Tenth Street
Bloomington, IN 47408
lewu@iu.edu

Carrie L. Ganote
Indiana University
2709 E. Tenth Street
Bloomington, IN 47408
cganote@iu.edu

Thomas Doak
Indiana University
2709 E. Tenth Street
Bloomington, IN 47408
tdoak@indiana.edu

Philip D. Blood
Pittsburgh Supercomputing Center
300 S. Craig Street
Pittsburgh, PA 15213
blood@psc.edu

Matthew Vaughn
Texas Advanced Computing Center
10100 Burnet Road
Austin, Texas 78758
vaughn@tacc.utexas.edu

ABSTRACT

The National Center for Genome Analysis Support (NCGAS) is a response to the concern that NSF-funded life scientists were underutilizing the national cyberinfrastructure. NCGAS is a multi-institutional service center that provides computational resources, specialized systems support to both the end-user and systems administrators, and most importantly scientific consultations to domain scientists unfamiliar with next generation DNA sequence data analysis. NCGAS is a partnership between Indiana University Pervasive Technology Institute, Texas Advanced Computing Center, San Diego Supercomputing Center, and the Pittsburgh Supercomputing Center. NCGAS provides hardened bioinformatic applications, user support on all aspects of a user's data analysis including data management, systems usage, bioinformatics, and biostatistics related issues.

Categories and Subject Descriptors

K.6.3 [**Software Management**]: Software maintenance; J.3 [**Computer Applications**]: Life And Medical Sciences—*Biology and genetics*

Keywords

NCGAS, Bioinformatics, Genome Analysis, Sequence Assembly

1. INTRODUCTION

The National Center for Genome Analysis Support (NCGAS) [3] is a response to the concern that NSF-funded life scientists were under utilizing the national cyberinfrastructure. To understand why this might be, you need to understand what these scientists need to compute. Biologists take observations on the physical world and then infer the state of unobservable entities or conditions via computation using models based on prior knowledge. In general, biologists have been in the forefront of mathematically rigorous quantitation since the dawn of modern science. It is not inappropriate to view the field of statistics as an extension of the theoretical underpinnings of biology and the early social sciences. But, within the last few decades new families of analytic instruments have been developed that can run simple biological experiments in a massively parallel fashion. Arguably the most well-known of these are the so called "next generation" DNA sequencers. Yet there are also "third generation" sequencers such as those from PacBio and potentially Oxford, hybrid mass spectrometers, and NMR instruments that are also opening new options for biological research. All of these instruments together have given rise to the omics fields; genomics, transcriptomics, proteomics, metagenomics, metabolomics, and recently multiomics, to name a few [4].

Informatics-aligned biologists specialize in various laboratory techniques associated with one class of modern instrumentation such as a DNA sequencer or a mass spectrometer, or conducting studies associated with a particular class of data such as RNA sequences. Also, the processing and interpretation of the output files of these instruments is a complex field in its own right, and is sometimes taken as the operational definition of bioinformatics. Thus, there are investigators that will refer to themselves as genomicists and specialize in determining the genomic sequence of, for example, single celled eukaryotes, and there are proteomic mass spectrometrists who understand protein sample preparation and tandem mass spectrometry, and there are bioinformaticists who would analyze the data from either or sometimes both of these researchers. But, the majority of life scientists do not do any of these things. Instead most biologists specialize in either some taxa of living organism or a biological process, or more likely both. Thus you will meet an ornithologist with an interest in behavioral ecology, or a physician-scientist with an interest in solid breast cancers. Either of these latter two researchers may benefit from this or that omic study, but only secondarily. Such a study would not be the center of their professional research. Instead, given the life science's long tradition of collaborative work, they will partner with the required specialists, or they will approach a "core" facility that specializes in one or a few omic techniques. So although these domain scientists increasingly need to accomplish omics science, they are not trained in the informatics disciplines associated with modern instrumentation, nor with using Unix-based high performance computing systems and further this training would be of little professional value to them; these skills are secondary to omic studies, which are secondary to their field of professional interest.

On September 15, 2011, Indiana University (IU) received three years of support to establish the National Center for Genome Analysis Support [3] to help NSF-funded life scientists leverage the national cyberinfrastructure to support their research.

In its first year and a half NCGAS has brought online a large-RAM computational cluster, recruited 37 NSF-funded genomics projects to use the resource, and processed 55,680 computational jobs representing a total of 16.4 core years of computing. NCGAS also laid the framework for creating a truly national-scale center supporting genomics research. By coordinating effort between multiple supercomputing centers, NCGAS is creating a service-oriented computational infrastructure - one that is designed to be approachable by end-users unaccustomed to using traditional supercomputing resources. In this paper we describe the organization of the center, the services we offer, and how NCGAS leverages XSEDE resources in support of the life science community.

2. CENTER ORGANIZATION

NCGAS is a multi-institutional center administered through Indiana University, but belonging to all the partner institutions; IU PTI [2], TACC [8], SDSC [7] and PSC [5]. Although all member institutions continue to support genomic science projects as they have in the past, by partnering in NCGAS the individual institutions are able to better leverage hardware and staff to meet research needs. For example, NCGAS bioinformatics and genomics staff at Indiana University can be referred trouble/help tickets from TACC should the need arise. Likewise, the computational tasks of NCGAS projects at IU that exceed the 0.5 TB RAM available on IU's Mason computer cluster can be easily transferred to either TACC or PSC. The synergistic benefits of such inter-institutional coordination can be seen from events such as the NCGAS co-hosted Daphnia Genomics Jamboree. At this gathering, dozens of scientists from across the US and Europe spent five days accelerating the completion of the Daphnia manga genome. NCGAS-supported staff from both TACC and IU gave technical presentations early in the Jamboree, before participants broke into small teams and used NCGAS clusters to perform their analyses.

2.1 Project Recruitment

NCGAS support is intended for NSF funded omics research projects that require large RAM clusters. Typically this includes genome assembly, transcriptomics, metagenomics and similar studies, but consideration will be given to any NSF funded life science project. Although NCGAS will gladly support any eligible project, staff have been particularly interested in supporting researchers in NSF Experimental Program to Stimulate Competitive Research (EPSCoR) states, and those at smaller institutions who do not have as easy access to large-scale computational or bioinformatic resources. Several instances have been identified where researchers at smaller institutions have installed modern sequencing equipment, but find that certain classes of experiments, particularly RNA-sequencing [12] on non-model organisms and metagenomic [17, 21] or metatranscriptomic [18] studies cannot be undertaken due to the lack of large RAM computational resources. The de novo assembly components of these studies can be conducted using NCGAS resources, while the remainder of the researchers' workflow is done on their own computers.

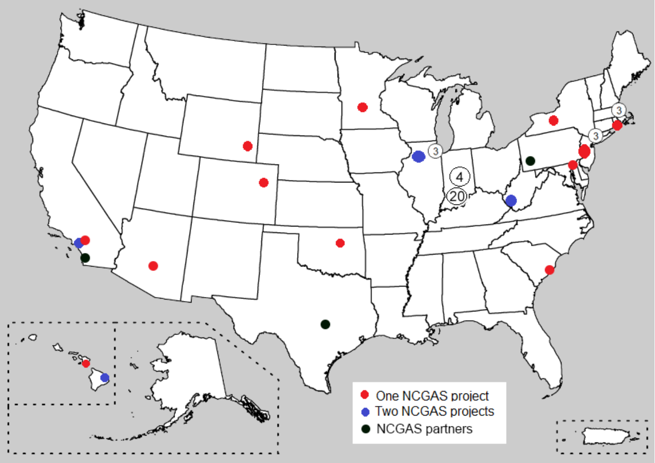

Figure 1: Map of projects: Updated to March 2013 values, and to include all three partners.

NCGAS has attempted to recruit projects needing support by attending and exhibiting at conferences that attract NSF funded life scientists who do not specialize in genomic sciences. This has worked well for identifying domain biologists for which genomics assembly is outside their primary research interest. In total, NCGAS staff spoke or exhibited at 18 meetings in the first year and a half of the center. Figure 1 shows the approximate location of the 50 NCGAS supported projects as of this writing.

2.2 Center Workflow

Typically NCGAS starts a relationship with a project when the primary investigator of the project, or his or her delegate, submits an allocation request at ncgas.org. NCGAS staff follow up the allocation request with a call or e-mail conversation to determine what level of support the project needs. This support typically falls along a gradient from projects that need nothing but access to computational resources (and perhaps a modest amount of technical support with the idiosyncrasies of the computational platform) at one end and projects where the researcher has no idea how to proceed with data analysis at the other. In the latter case, staff conducts a "Science Call" where Ph.D.-level genomicists and bioinformaticians, along with NCGAS data analysts and computational support staff share a conference call with the researchers to determine what tasks need to be done, and who will do them. Whenever possible, center staff try to train individuals at the research laboratories to run the repetitive portions of their analysis themselves.

3. SERVICES PROVIDED

Access to NCGAS computational and consulting services is awarded to qualified genomics research projects through an allocation process. To request an allocation, researchers need to submit the NCGAS Allocations Request form, where they are asked to provide information about the project's principal investigator, the related NSF award number, and the personal information about each individual who will need an account on NCGAS systems. Allocation are reviewed and awarded within one business day of the request. NCGAS provides access to large RAM clusters, technical support in using the clusters, bioinformatic support in determining the most appropriate method to analyze a given data set, and actual assistance in running the computation. NCGAS particularly supports the analysis of next-generation sequencer output in the following categories:

- De novo assembly, which does not use a reference genome and requires that each read be compared to every other to find overlaps, usually in transcriptomic studies.
- Metagenomic projects, which simultaneously sequence the combined genomes in an environmental sample, such as ocean water or the human mouth.
- Resequencing, where a closely related genome has already been completely sequenced and assembled.

3.1 Bioinformatic Software Support

NCGAS installs and supports bioinformatic software. Frequently community-developed tools can be challenging to correctly install on HPC clusters. Therefore NCGAS staff are tasked with installing and verifying key applications used by supported projects. Table 1 lists the genomics assembly and bioinformatic applications NCGAS supports by system.

Additionally, NCGAS works with computer scientists interested in improving the performance characteristics of bioinformatic software. For example NCGAS, partnered with The Broad Institute [1], the IU Pervasive Technology Institute [2], and ZIH [9] in Dresden to optimize the best-in-class RNA-sequence assembly application Trinity for use on large RAM HPC systems [6, 15].

3.2 NCGAS at Indiana University

Indiana University is the hub of NCGAS activity. IU provides the Mason cluster, center staff, and administrative support. Additionally, through a collaborative effort between IU and Penguin Computing [1], NCGAS services and support are made available to non-NSF funded researchers on the Penguin On-Demand (POD) HPC Cloud Service. POD has installed within the IU data center Rockhopper, a large-memory supercomputing cloud appliance hosted by IU, and equipped with genome analysis and bioinformatics software. Researchers without NSF support can purchase computing time from Penguin Computing, and receive access to NCGAS hardened applications, and technical and bioinformatic support from IU-sponsored NCGAS staff.

In general, NCGAS at IU provides the following consulting service:

1. Support for users of genome analysis software.
2. Storage of submitted data sets and results for at least one year following analysis.
3. A repository of open source genome analysis software, including hardened, tuned, and optimized versions of particularly important open source software.
4. Support for use of open source genome analysis software, including extended, in-depth consulting for sequence analysis, and tutorials and presentations about using NCGAS services.
5. Access to a custom Galaxy instance.

NCGAS at IU has established several Galaxy web portals [10, 13, 14], shown as Table 2, to allow researchers to use either Mason or Rockhopper with a familiar web interface. For example, the Galaxy @ IU portal employs IU's Mason cluster for compute services and the IU Data Capacitor for project storage [16], and is hosted on IU's Quarry Gateway Web Services Hosting System. Access requires an IU Network ID.

Galaxy instance	Required affiliation	Additional information
Galaxy @ IU	IU students, faculty, and staff	Provides institutional support for life science researchers.
Galaxy @ NCGAS	Researchers with NCGAS allocations	Access requires an NCGAS allocation
Galaxy @ Rockhopper on POD IU	Researchers on the POD IU (Rockhopper) cluster	Rockhopper provides fee-based, on-demand cloud computing service. Access requires purchase of on-demand services.

Table 2: Galaxy services at IU. (The Galaxy project is supported in part by the NSF, the National Human Genome Research Institute (NHGRI), and PSU's Huck Institutes of the Life Sciences.)

The interaction between Indiana University NCGAS staff and the user community is classified into three categories: short-term consultations, long-term consultations, and supported projects. Short-term consultations take less than four hours of staff time and typically center on resolving a simple technical question, or advising a user on how to proceed. Long-term consultations require more than four hours of effort and can be either technical or scientific. The former usually revolve around complex technical issues that exceed the reasonable understanding of a domain scientist. These include requests to install software packages with complex dependencies or to troubleshoot error messages received from failed jobs. It is not uncommon for these consultations to require NCGAS staff to interact with the user, the systems administrators, and the software community that developed the failing software. Scientific long-term consultations are primarily the domain of Drs. Doak and LeDuc. As a genomicist and biometrician respectively, they can either advise domain scientists on appropriate methodologies for processing and analyzing data, or help put the requesting scientist in contact with a knowledgeable expert. Lastly there are supported programs. This category is used to describe separate research initiatives or individual grant recipients who

[1] http://www.penguincomputing.com/

have asked for allocations on NCGAS resources. In the first NCGAS project year staff reported 480 short-term consultations, 22 long-term consultations, 25 NSF-funded supported programs, and 14 non-NSF-funded supported programs.

3.3 NCGAS at TACC

A key aspect of enabling and enhancing accessibility of high performance computing systems to biologists engaged in genomics research is ensuring that the codes they use can run to completion quickly and with a reasonable degree of efficiency. Unfortunately, many genomics community codes are not poised to take advantage of advances in microprocessor technology, and in fact, as the size of input data grows, they will run progressively more slowly. TACC graduate assistant Manoj Dhanpal has led a year-long effort to explore the feasibility of adapting over a dozen community bioinformatics codes for acceleration on both traditional multicore x86 and the new many core Xeon Phi. These have included BWA (Burrows-Wheeler Aligner) [2], components of Trinity [6], NCBI BLAST [3], Bowtie [4], Samtools [5], and several C-based Bioconductor components [6]. Dhanpal and collaborators have compiled these codes using the Intel C compiler and run benchmarks on these codes using tools such Perfexpert [11] and Valgrind [19] to identify several optimzation opportunities. Along the way, several issues involving linker mechanics and lack of SSE intrinsics (used by many genomics codes) have been brought to the attention of the Intel compiler development team. We envision that this fundamental work will pave the way for early adopters of the Intel MIC architecture to make performance gains on a variety of genomics codes.

3.4 NCGAS at PSC

The Pittsburgh Supercomputing Center (PSC) is the newest NCGAS partner, joining in January 2013. The PSC has a long history of supporting bioinformatics through its National Resource for Biomedical Supercomputing, formerly the Biomedical Initiative, established in 1987 and funded by the NIH. In 2010, the NSF funded the PSC to deploy and support the largest shared memory machine in the world. This resource, named "Blacklight", is an SGI Altix UV 1000 with two 16 TB partitions of cache-coherent shared memory with 2048 cores each. A single application on Blacklight can therefore use up to 16 TB of shared RAM. Since the deployment of Blacklight, the PSC has supported a wide range of genomics applications, specializing in the most challenging de novo assembly problems, which may require hundreds of gigabytes or even terabytes of shared memory.

For instance, PSC recently teamed up with researchers from Cornell to create the Non-human Primate Reference Transcriptome Resource (http://nhprtr.org/) [20], which so far has required the de novo assembly of more than 20 primate transcriptomes from 13 different species. Using the Trinity code, each of these assemblies required up to 1.5 TB of RAM. PSC also worked with researchers at Oklahoma State to assemble 300 gigabases of soil metagenome data, requiring 3.5 TB of RAM, to look for genes that may assist in developing new biofuels. In all of these activities, the PSC shares the goal of NCGAS not just to support individual research activities, but to enable a community of researchers adept at using high-end computing resources to tackle the most difficult problems in genomics. This is being accomplished in part by developing close ties with both researchers and developers of key community codes, like Trinity and Galaxy, and enabling these codes, which researchers are already using, to seamlessly utilize Blacklight and other advanced resources on XSEDE.

4. NCGAS AND XSEDE

NCGAS is both an XSEDE user and provider of a Tier 2 resource. NCGAS makes 25% of Mason available to individual NSF-funded XSEDE genomics projects using the XSEDE allocation system. Mason supports both small startup allocations of 15,000 SU and larger XRAC allocations of 300,000 SU. This allows end-users familiar with the XSEDE allocation process to gain access to the hardened bioinformatic applications installed on Mason, in addition to those installed on other XSEDE systems.

As an XSEDE user NCGAS relies on XSEDE resources to facilitate workflows across partner institutions. For example, NCGAS promotes a "compute in place" strategy where researchers load their data once on to the Data Capacitor's WAN (DCwan), a distributed Lustre file system. The DCwan is an XSEDE resource [11] that is mounted on all NCGAS partner systems. This allows data files created by computation on one system to be immediately available on any other.

NCGAS has an XSEDE startup allocation for the installation and benchmarking of bioinformatic applications across the various partner systems. This small allocation also serves as the prototype for a Community Allocation that is intended to allow researchers to perform all steps in a biological workflow without needing to move their data. For example, in a typical RNA-sequencing experiment there will be a large de novo assembly task using Trinity to generate gene models [6][10]. This task needs to be done on a large RAM cluster. But once the gene models are generated, each of the 20 to 50 thousand gene models will need to be annotated using Blast, which is best run on traditional clusters. In the NCGAS workflow, the assembly might be done on Mason while the annotation is done on systems at TACC – but all data movement is transparent to the user.

In addition, NCGAS is actively working to simplify access to partner resources. For example, NCGAS has received an ESCC startup allocation to assist in identifying a middleware layer that can be used to manage the communication between the NCGAS Galaxy installation and all of the partner systems. It is imagined that this Galaxy installation can be configured such that the end user would be completely concealed from, not only the complications of moving data between systems, but also the cumbersome and error-prone command line interface.

5. FUTURE DIRECTIONS

5.1 Expanding the User Base

Next generation sequence-based research is inherently episodic. Domain scientists determine that they have a need for a study, conduct it, and then briefly need NCGAS resources. For example, a phytoplankton researcher may be interested

[2] http://bio-bwa.sourceforge.net/

[3] http://blast.ncbi.nlm.nih.gov/

[4] http://bowtie-bio.sourceforge.net/

[5] http://samtools.sourceforge.net/

[6] http://www.bioconductor.org/

in phosphorus processing in seawater, and determine that a eukaryotic metatranscriptomic study comparing three different locations and two different depths of water would be useful in understanding the problem. The researcher secures the funding, designs the study, collects the samples (which in this case would require an ocean going research vessel fully crewed etc.), processes them and sends them for sequencing. It can easily take a calendar year from the conception of the study until the sequence data is available for analysis. For a few weeks the researchers attend to analyzing the sequence data, and then they retire to draw meaningful conclusions from the experimental results. It may be well over another year before that researcher needs to analyze more sequence data. The need for NCGAS resources, from the point of view of the researcher, is short but intense; this creates an environment where NCGAS will need to continue its outreach services for the foreseeable future. The Center will need to let researchers know about the availability of support early in the design of studies, but for individual experiments it may be months before the services are needed. Thus, in order to maintain a near constant utilization of resources NCGAS will need the largest stable of research projects that can be assembled. Therefore it is expected that NCGAS will continue having a presence at domain specific conferences and other outreach venues for the foreseeable future.

5.2 Future Services

NCGAS has two approaches going forward to assist NSF-funded life scientists in using national cyberinfrastructure to analyze omics data. First NCGAS is working towards an XSEDE Community allocation which can be shared with those users comfortable with the launching jobs on existing XSEDE resources. This allocation will allow NCGAS users to bypass the lengthy XSEDE allocation process needed to receive an allocation for a request that is itself only a little larger than a startup allocation. Further, with NCGAS staff available to provide support and guidance, users can approach XSEDE resources with less investment of time and needing less training.

Perhaps more importantly, NCGAS is planning on deploying a Galaxy instance that presents all the hardened bioinformatic tools needed to complete commoditized data analysis workflows, and that can seamlessly submit jobs across any NCGAS partner resource. Figure 2 shows how this will be done. Issues related to data movement are handled below Galaxy by the distributed file system. All that is needed is an appropriate middleware for communicating jobs from the web portal to the various systems, and job status and exit conditions back. As far as the web portal is concerned all data and compute is local, so the only modification needed to the Galaxy system is a modified job runner that interfaces with the selected middleware. This Galaxy instance will be an extension of the existing NCGAS Galaxy and as such it will only provide those tools that are needed for workflows that are relevant to next generation sequence assembly, and closely related tasks. Since the Galaxy instance will be used for a limited set of bioinformatic tasks that require a well defined set of applications that tend to run long walltime jobs on a reasonable amount of data in a unidirectional manner, it is expected that this arrangement will give a satisfactory user experience.

6. ACKNOWLEDGEMENTS

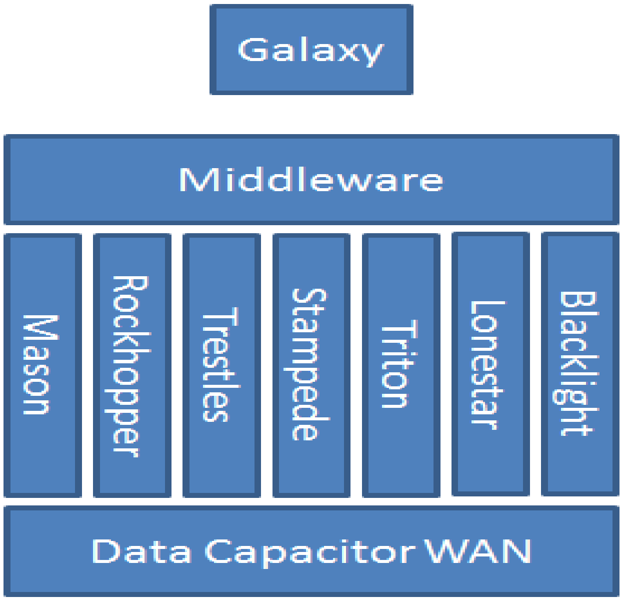

Figure 2: A schematic showing how a custom Galaxy can use a middleware layer and a distributed file system to support genomic analysis across a number of NCGAS supported systems.

This work was supported in part by NCGAS funded by NSF under award No. 1062432. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

7. ADDITIONAL AUTHORS

Additional authors: William K. Barnett (Indiana University, email: barnettw@iu.edu)

8. REFERENCES

[1] Broad Institute . http://www.broadinstitute.org.
[2] Indiana University Pervasive Technology Institute. http://rc.uits.indiana.edu.
[3] National Center for Genome Analysis Support. http://www.ncgas.org/.
[4] Omics Gateway. http://www.nature.com/omics/about/index.html.
[5] Pittsburgh Supercomputing Center. http://www.psc.edu/.
[6] RNA-Seq De novo Assembly Using Trinity. http://trinityrnaseq.sourceforge.net.
[7] San Diego Supercomputer Center. http://www.sdsc.edu/.
[8] Texas Advanced Computing Center. http://www.tacc.utexas.edu/.
[9] The Center For Information Services And High Performance Computing. http://www.tu-dresden.de/die_tu_dresden/zentrale_einrichtungen/zih.
[10] D. Blankenberg, G. Von Kuster, N. Coraor, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko, and J. Taylor. Galaxy: a web-based genome analysis tool for experimentalists. *Current protocols in*

molecular biology / edited by Frederick M. Ausubel ... [et al.], Chapter 19, Jan. 2010.

[11] M. Burtscher, B.-D. Kim, J. Diamond, J. McCalpin, L. Koesterke, and J. Browne. Perfexpert: An easy-to-use performance diagnosis tool for hpc applications. In *Proceedings of the 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '10, pages 1–11, Washington, DC, USA, 2010. IEEE Computer Society.

[12] R. Ekblom and J. Galindo. Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, 107(1):1–15, Jul 2011.

[13] B. Giardine, C. Riemer, R. C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, W. Miller, W. J. Kent, and A. Nekrutenko. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.*, 15(10):1451–1455, Oct 2005.

[14] J. Goecks, A. Nekrutenko, J. Taylor, and T. G. Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8):R86, 2010.

[15] R. Henschel, M. Lieber, L.-S. Wu, P. M. Nista, B. J. Haas, and R. D. LeDuc. Trinity rna-seq assembler performance optimization. In *Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the eXtreme to the campus and beyond*, XSEDE '12, pages 45:1–45:8, New York, NY, USA, 2012. ACM.

[16] R. Henschel, S. Simms, D. Hancock, S. Michael, T. Johnson, N. Heald, T. William, D. Berry, M. Allen, R. Knepper, M. Davy, M. Link, and C. A. Stewart. Demonstrating lustre over a 100gbps wide area network of 3,500km. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, SC '12, pages 6:1–6:8, Los Alamitos, CA, USA, 2012. IEEE Computer Society Press.

[17] P. Hugenholtz and G. W. Tyson. Microbiology: Metagenomics. *Nature*, 455(7212):481–483, Sep 2008.

[18] S. Mitra, P. Rupek, D. C. Richter, T. Urich, J. A. Gilbert, F. Meyer, A. Wilke, and D. H. Huson. Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG. *BMC Bioinformatics*, 12 Suppl 1:S21, 2011.

[19] N. Nethercote and J. Seward. Valgrind: a framework for heavyweight dynamic binary instrumentation. In *Proceedings of the 2007 ACM SIGPLAN conference on Programming language design and implementation*, PLDI '07, pages 89–100, New York, NY, USA, 2007. ACM.

[20] L. Pipes, S. Li, M. Bozinoski, R. Palermo, X. Peng, P. Blood, S. Kelly, J. M. Weiss, J. Thierry-Mieg, D. Thierry-Mieg, P. Zumbo, R. Chen, G. P. Schroth, C. E. Mason, and M. G. Katze. The non-human primate reference transcriptome resource (NHPRTR) for comparative functional genomics. *Nucleic Acids Res.*, 41(Database issue):D906–914, Jan 2013.

[21] R. Seshadri, S. A. Kravitz, L. Smarr, P. Gilna, and M. Frazier. Camera: A community resource for metagenomics. *PLoS Biol*, 5(3):e75, 03 2007.

Software	Mason (IU)	Rockhopper (IU)	Trestles (SDSC)	Stampede (TACC)	Lonestar (TACC)	Triton (SDAC)	Blacklight (PSC)
abyss	✓	✓	✓		✓		✓
allpathslg	✓	✓			✓		✓
amber			✓	✓	✓		✓
amos	✓				✓		
arachne	✓						
autodock				✓	✓		✓
beagle					✓		
beast			✓		✓	✓	
bedtools	✓			✓	✓		
bio3d	✓						✓
bioconductor	✓						
bioperl	✓	✓			✓		✓
biopython	✓						✓
bioscope							✓
bismark			✓	✓	✓		
bitseq	✓				✓		
blast		✓	✓		✓		
blat	✓			✓	✓	✓	✓
boost	✓		✓	✓	✓		✓
bowtie	✓	✓	✓	✓	✓	✓	✓
bwa	✓	✓		✓	✓	✓	✓
cafe	✓						
cd-hit	✓						
celera	✓	✓					
clustalw2	✓			✓	✓		
cufflinks	✓		✓		✓		
cytoscape	✓						
edena	✓						
euler	✓						
fastq			✓	✓			
fastx-toolkit	✓			✓	✓		
garli			✓		✓		
gatk	✓			✓	✓		
genomemapper	✓						
gmap	✓			✓	✓		
mlRho	✓						
mrbayes			✓	✓	✓		
mummer	✓		✓	✓	✓		
muscle					✓		
namd	✓		✓	✓	✓	✓	✓
newbler				✓			
ninja	✓						
novoalign	✓						
oases	✓			✓	✓		✓
phylip				✓	✓		
phyml					✓		
phyutility				✓	✓		
picard	✓				✓		✓
plink				✓	✓		
raxml	✓		✓	✓	✓		
samtools	✓		✓	✓	✓		✓
shore	✓						
smrt	✓						
soap				✓	✓		✓
soapdenovo	✓	✓	✓				✓
soaptrans				✓	✓		✓
sra-toolkit	✓			✓	✓		✓
tophat	✓		✓	✓	✓		
transabyss	✓						✓
trinityrnaseq	✓	✓			✓		✓
vcftools	✓						
velvet	✓		✓	✓	✓		✓

Table 1: NCGAS software support across partners as of the end of PY1. This table shows the level of shared support for bioinformatics software on the various NCGAS partner compute resources in PY1.