National Center for Genome Analysis Support
Indiana University Pervasive Technology Institute

XSEDE
Extreme Science and Engineering Discovery Environment

PERVASIVE TECHNOLOGY INSTITUTE
INDIANA UNIVERSITY

National Resources for Computationally Intensive Genomics

Carrie Ganote Barb Hallock

Genomics as a data-intensive science

Genomics has taken its place as a card-carrying member of Big Data science. Next Generation sequencing technologies have increased the throughput of sequencing while decreasing the cost dramatically (Figure 1). The plummeting cost has led to a proliferation of techniques to explore other phenomnena such as transcriptomics, methylation and protein binding.

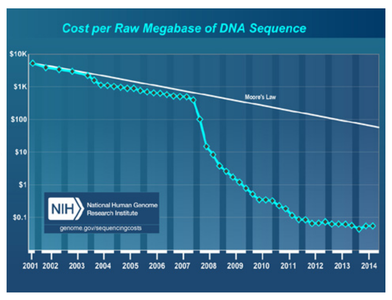


Figure 1. Cost decrease of DNA Sequencing (1)

Types of computational resouces

Basics

Genomics studies can involve a wide variety of hardware and software solutions that will lead to a completed analysis, but to really make the best use of your time and available resources, it is best to pick the best tool for each job. Not only are you going to need a computer to run analyses on, but you may need a way to get your data to that machine. You will want to eventually save the raw inputs as well as the finished product when it time to publish. The right software and parameters will make a potentially large difference in the outcome of the experiment. Use the available national cyberinfrastructure to its fullest to save yourself time, money, and manuscript revisions.

High Performance Computing (HPC)

This is what usually comes to mind when thinking about cyberinfrastructure. Often called "supercomputers", these large scale machines are designed to be used for jobs that a desktop can't handle. Supercomputers are clusters of individual computers that are wired together to allow very fast communication between each computer, which is called a node. Individual nodes will have many CPU cores and often a respectable amount of RAM. Their ability to communicate with each other allows a researcher to run a big job using many nodes at once, as though it were running on one single machine. Jobs that work well on HPC require a lot of memory, processing power, or disk space and cannot be broken down into smaller jobs.

High Throughput Computing (HTC)

There are problems in genomics that are not useful to throw at the biggest, fastest machine out there, but lend themselves very well to using hundreds to thousands of moderately sized computers. Approaches that work well for this might involve large input files where each line can be treated independently and does not depend on the value of any other input line. As long as the resources required for each separate input are small, this is a very scalable solution.

Data Storage and Movement

One of the most important considerations (and often an overlooked aspect of data analysis) is the handling of the data before, during and after an experiment. Large data may be generated at a remote facility and will need to be moved to the site where it will be analyzed. Input files will need to be archived in an accessible way to provide reproducabillity of your work. Careful management plans that include cleaning out junk files will utilize available space and lead to less stress over storage quotas.

Software and Savvy

Powerful hardware increases the potential work a researcher can accomplish, but without the appropriate tools and training these advantages may be lost. Bioinformatics software should be installed, maintained and supported to allow users to get the most out of the system. Environments range from command-line-only setups to fully web-based graphical interfaces and everywhere in between. Training, collaboration and consulting serve

Where to find computational resources

Your Institution

This should be the first place you look! Many academic and research organizations will have some kind of support for their own members. Colleges and Universities may give priority to their own faculty, staff and students and provide free access to high performance computing platforms. If they don't, make some noise and express your needs. Reach out to your IT or Computer Science department. It's possible that students are looking for a project just like yours to use as a pilot project!

NCGAS

NCGAS is an NSF-funded project that provides computing power and bioinformatics expertise to the genomics research community. We are happy to offer advice and talk about our experience with anyone; long-term support and and computational resources are typically provided to other NSF-funded projects or projects that fall under the purview of the NSF. We also write letters of commitment of our resources if you are currently writing a proposal.

XSEDE

Extreme Science and Engineering Discovery Environment (XSEDE) is a great resource for researchers who need to dabble in HPC solutions (2). There are many ways to be involved with XSEDE - the best way to start is to contact your Campus Champion. Startup allocations provide the nest tier of resources if your needs are small or if you need to benchmark your workflow to apply for a full research allocation.

iPlant

The iPlant Collaborative (3) provides many tools for analyzing genomic data. The Discovery Environment and DNA Subway provide a graphical user interface for point-and-click job submission.
Atmosphere is iPlant's cloud interface where users can set up a Linux environment of their choice, with pre-loaded software and an IP address for hosting genome browsers.

OSG

The Open Science Grid (4) provides high-throughput solutions to researchers. Institutions which join the OSG donate their opportunistic - otherwise unused - compute time to the grid. In exchange, users from the institution can send jobs to run on the OSG when their own machines are busy. This allows for greater utilization of these resources and shorter wait times for the researcher. The types of machines that are on the grid are widely varied, so jobs that fit a lower denominator in terms of compute power will benefit the most.This might be a good solution for certain sets of problems that are embarassingly parallel, such as BLAST+ (5).

Galaxy

Galaxy (6-8) is a web portal for bioinformatics tools that can be hosted anywhere. NCGAS maintains its own Galaxy instance as part of its service suite, but Galaxy Main is a public resource and a great way to do heavy computation without needing a degree in Computer Science.

CIPRES

CIPRES (9), or the Cyberinfrastructure for Phylogenetic Research, is a web portal that provides tools for phylogenetic analysis. Tools for tree building and estimation are provided through a graphical interface that ties to XSEDE hardware.

Amazon Education Grant

Amazon offers credit for several of its popular services (terms):

- Amazon Simple Storage Service
- Amazon CloudFront
- Amazon Simple Queue Service
- Amazon Elastic Compute Cloud
- Amazon SimpleDB Service
- Amazon Relational Database Service
- Amazon ElasticMapReduce

The credit provided by this grant is charged at the usual rates listed through Amazon Web Services. This allows an awarded researcher flexibility in what resources he or she can use at a given time.

Department of Energy INCITE Award

Open to a broad audience, the INCITE award provides very large allocations of compute time on the Titan Supercomputer to high risk + high reward projects. The kind of work that wins an award like this one would likely be a real game-changer, have high impact, and have far-reaching implications.

Conclusion

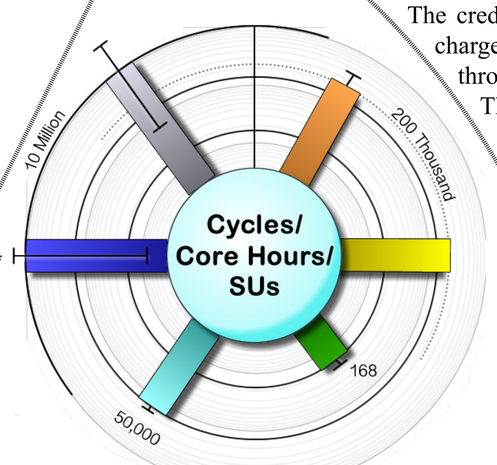
Expensive hardware and the costs of maintaining a full time system administrator limit the availablility of appropriate tools to the genomicist today. Many publicly-funded options are available to take advantage of. Many of these resources take very little time to sign up for, or otherwise will grant an allocation that can cover all of your computational needs.

For more information, visit ncgas.org.

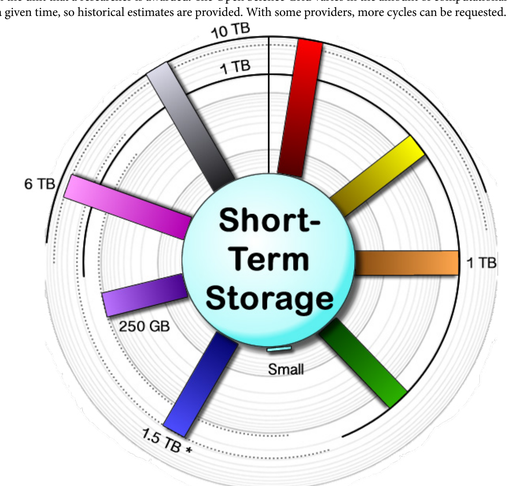
Breakdown of Resources

Funding/ Restrictions	Registration Process	Deadlines	Allocation Period	Renewable
NSF funded or fundable	Medium/ Registration	None	2 Years	Yes
Faculty or Post-Doc	Medium/ Allocation	None	1 Year	No
Faculty or Post-Doc	Difficult/ Allocation	jan.15, Apr.15, Jul. 15, Oct. 15	1 Year	Yes
No Restriction	Easy/ Registration	None	No Restriction	-
No Restriction	Easy/ Registration	None	No Restriction	-
No Restriction	Medium/ Registration	None	No Restriction	-
No Restriction	Easy/ Registration	None	No Restriction	-
Accredited University	Medium/ Allocation	March 31, June 30, Sept.30, and Dec. 31.	2 Years	Yes
No Restriction	Difficult/ Allocation	mid-April to the end of June	1 Year	No

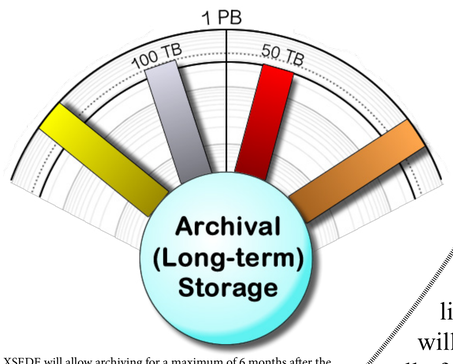
Table 1. Restrictions on the availability of each resource. Some resources will only allow certain funding sources for the projects they support, or the requestor must belong to an academic institution. Some resources require an allocation request which may be reviewed by a committee. Others follow a simple registration process.



Core hours, Service Units or generally "cycles" are the amount of computation that can be done by one CPU in one hour, and are often the unit that a researcher is awarded. The Open Science Grid varies in the amount of computational resource available at a given time, so historical estimates are provided. With some providers, more cycles can be requested.

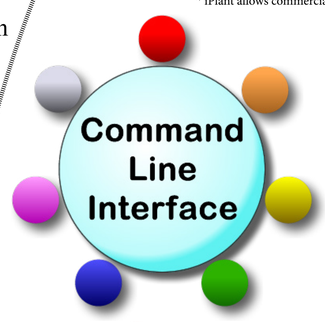


Short term storage is the amount of space generally available for the duration of the job. The Open Science Grid varies in available space; the login node provides 1.5TB and 2GB is available for software binaries.



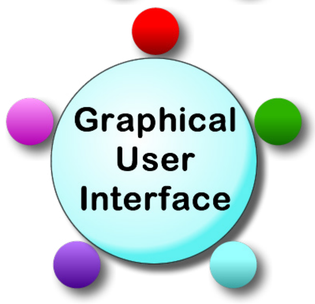
XSEDE will allow archiving for a maximum of 6 months after the completion of the XSEDE allocation. NCGAS will hold data and make it available for publishing through IUScholarWorks.

Some resources allow their services to be used for commercial, proprietary and for-profit work.
* iPlant allows commercial users at its discretion.



- NCGAS
- XSEDE Startup
- XSEDE Research Allocation
- iPlant
- OSG
- Amazon Web Services
- DOE INCITE

Many HPC and HTC resources provide access to the command line. Access may be granted via an SSH shell, or in the case of iPlant, an Atmosphere instance that can run the terminal program. Command line access and graphical user interface access are not mutually exclusive.



- NCGAS
- iPlant
- CIPRES
- Galaxy Main
- Amazon Cloud - A user interface may be used as part of Amazon Cloud, though this is not strictly a feature of Amazon.

A graphical user interface may be provided to the user for an easy way to run complex, time-intensive analyses without the need for a shell or command line interface. Some resources provide both.

Many resource providers are funded by the US government, and as such, they restrict access to grants and allocations to US citizens. Most will still allow collaborators on projects with a US institution to apply. These resources are open worldwide.
* iPlant allows international use at its discretion

Walltime	Cores per Job	(GB) RAM per Node	Concurrent Jobs
14 Days	32	512	8
7 Days*	16,000*	1000+*	50*
7 Days*	16,000*	1000+*	50*
24 Days	16	128	No Restriction
7 Days	Unknown	Unknown	No Restriction
**	**	**	No Restriction
36-48 Hours	6 to 16	32	6
No Restriction	32	224	No Restriction
1 Day	100 aprun processes/batch	32 + 6	2

Table 2. Technical Details. Resources will vary in the specifications and limitations of their hardware.
* XSEDE is a collection of different institutions, each with different hardware choices. Details will depend on what machine you run on. Listed here are the upper bounds for these values.
** The Open Science Grid uses a heterogenous set of resources and thus limits depend on the available computers. The more you ask for, the less likely a worker node will be available to fulfil your request.

References

(1) Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcosts. Accessed Jan 9, 2015
(2) Towns, J. et al. (2014). "XSEDE: Accelerating Scientific Discovery", Comp. in Sci. & Eng., vol.16, no. 5, pp. 62-74, Sept.-Oct. 2014, doi:10.1109/MCSE.2014.80
(3) Goff, S.A. et al. (2011). "The iPlant Collaborative: Cyberinfrastructure for Plant Biology," Frontiers in Plant Science 2, doi: 10.3389/fpls.2011.00034.
(4) Pordes, R. et al. (2007). "The Open Science Grid", J. Phys. Conf. Ser. 78, 012057.doi:10.1088/1742-6596/78/1/012057.
(5) Camacho C. et al. (2008) "BLAST+: architecture and applications." BMC Bioinformatics 10:421.
(6) Goecks, J. et al. (2010). "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences". Genome Biol. 2010 Aug 25;11(8):R86.
(7) Blankenberg, D. et al. (2010). "Galaxy: a web-based genome analysis tool for experimentalists". Current Protocols in Molecular Biology. 2010 Jan; Chapter 19:Unit 19.10.1-21.
(8) Giardine, B. et al. (2005). "Galaxy: a platform for interactive large-scale genome analysis." Genome Research. 2005 Oct; 15(10):1451-5.
(9) Miller, M.A. et al. (2010) "Creating the CIPRES Science Gateway for inference of large phylogenetic trees" in Proc. of the Gateway Comp. Env. Workshop (GCE), 14 Nov. 2010, pp 1 - 8.

Informational Pages:
https://pods.iplantcollaborative.org/wiki/display/atmman/Allocation+Policies
https://proposals.doeleadershipcomputing.org/allocations/calls/incite2015
http://www.doeleadershipcomputing.org/guide-to-hpc/
https://www.olcf.ornl.gov/support/system-user-guides/titan-user-guide/
http://aws.amazon.com/ec2/instance-types/
http://aws.amazon.com/education/terms/
http://aws.amazon.com/education/faq/
http://aws.amazon.com/grants/

Legal
This material is based upon work supported by the National Science Foundation under Grant No. ABI-1062432, Craig Stewart, PI. William Barnett, Matthew Hahn, and Michael Lynch, co-PIs. This work was supported in part by the Lilly Endowment, Inc. and the Indiana University Pervasive Technology Institute. Any opinions presented here are those of the presenter(s) and do not necessarily represent the opinions of the National Science Foundation or any other funding agencies.
This document is released under the Creative Commons Attribution 3.0 Unported license (http://creativecommons.org/licenses/by/3.0/). This license includes the following terms: You are free to share – to copy, distribute and transmit the work and to remix – to adapt the work under the following conditions: attribution – you must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work). For any reuse or distribution, you must make clear to others the license terms of this work.