HPC Resources

- **XSEDE** - The Extreme Science and Engineering Discovery Environment (https://www.xsede.org/), If you are a US-based researcher and currently use, or want to use, advanced research computing resources and services, XSEDE can help (an NSF-funded resource).
- **Jetstream** – A XSEDE cloud computing resource (https://jetstream-cloud.org/) funded by NSF
- **Virtual Machines (VM) -** An emulation of an actual computer which uses a set virtual memory and disk spaces, can host a completely different operating system. In cloud computing systems (like Jetstream and AWS), you can spin up a virtual machines.
- **High Performance Computing/Clusters (HPC)** - High-performance computing (HPC) is the use of super computers and parallel processing techniques for running advanced application programs efficiently and quickly.
- **Science Gateways (or Portals) -** An interface designed specifically to support a particular type of scientific research, with an emphasis on supporting the entire scientific process from start to finish (https://kb.iu.edu/d/auwv)
- **Cloud computing –** cloud computing is the delivery of computing services—servers, storage, databases, networking, software, analytics, and more—over the Internet. XSEDE offers Jetstream for researchers, and there is Amazon Web Services (AWS).
- **Sysadmin -** System administrator, the people who manage the machines and often set up the job schedulers. You should make friends with these people ^_^.

Anatomy of a Cluster

- **Login nodes –** Also called head nodes (often start with h when you log on - i.e. h2). When you ssh into a cluster/server. User process up to a limited time or memory (RAM) are allowed here. The limitation is based on the cluster. At IU, login nodes are limited to 20 minutes of CPU time.
- **Compute nodes -** If your application requires more than limited resources on login node, submit a batch job or an interactive session using the TORQUE qsub command or SLURM sbatch command. Often start with c when you log onto a computer node for an interactive job - i.e. c294.
- **Job schedulers-** Resource management system for submitting and controlling jobs on supercomputers, clusters, and grids. For more information: TORQUE - https://kb.iu.edu/d/avmy; SLURM - https://www.psc.edu/bridges/user-guide/running-jobs
- **Filesystems-** Controls how data is stored and retrieved. Without a file system, information placed in a storage medium would be one large body of data with no way to tell where one piece of information stops and the next begins (https://en.wikipedia.org/wiki/File_system). At IU we user lustre filesystems - https://kb.iu.edu/d/ayfh
- **Modules-** Some software is already installed and available as modules. A modulefile contains the user environment changes required for accessing the application. For more information on how to look for module files and add a module file to the user environment: https://kb.iu.edu/d/aosi. To request a module https://pti.iu.edu/contact/module/index.html
- T**ape versus Disk Storage -** Tape backup provides the ability to copy data packages from a hard drive to a tape cartridge for storage, backup and recovery purposes in the event of a

computer crash or other failure. Conversely, disk storage maintains data and other information on a hard drive fork rapid access for computation. The difference between the two includes speed, capacity and cost.

Misc Cluster Terms

- **CLI** - Command Line Interface, users can interact with devices and programs only through text input.
- **GUI -** Graphical User Interface, users can interact with devices and programs through graphical icons.
- **GPU –** Graphics processing unit- their highly parallel structure makes them more efficient than general-purpose CPUs for algorithms where the processing of large blocks of data is done in parallel.
- **IO operations-** Input/Output operations, where information is read from or copied to RAM or from disk. These are of most concern when there are lots of small files being produced or lots of small files being referenced - each time, the processor has to go find it and that adds up! Can be a limiting step in some computations.
- **RAMDisk -** portion of your system memory (RAM) is used as a disk drive (disk space), this increases the efficiency of IO operations
- **Stderr –** Standard error, generally saved as <job name>.e<job id>.
- **Stdout –** standard output, generally saved as <job name>.o<job id>.
- **Nohup –** No hangup signal, prevents the command from being aborted automatically when you log out or exit the shell (https://www.computerhope.com/unix/unohup.htm). Similar to *screen*

DE Analyses

- **p-value** - the chance a difference was found by chance for individual statistical tests
- **fold change -** a measure of how much a quantity changes based on an initial value (control). Usually expressed as the log2 value of the amount of change (30/60 = change of 2, log2 of 2 = fold change of 1).
- **FDR -** means of controlling for multiple tests in differential expression. Less conservative than Bonferroni correction, reports a q value.
- **q-value -** adjusted p value for multiple tests. The % of significant results that are found by chance. For example, if you have a p value of 0.05 for 100,000 results and 20,000 were determined positive results, 5,000 of them could be "false positives" (5% of 100,000). If you have a q value of 0.05 for 100,000 results and 20,000 were determined to be positive, 1,000 of them could be "false positives" (5% of 20,000).
- **FPKM -** Fragments Per Kilobase Million (https://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/)
- **TPM -** Transcripts Per Kilobase Million (https://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/)
- **RPKM -** Reads Per Kilobase Million (https://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/)
- **homeologs** - A special case of paralogy resulting from polyploidy. Created through chromosome duplication events https://genomevolution.org/wiki/index.php/Homeolog

NCGAS Resources

- **NCGAS mainpage -** https://ncgas.org
- **NCGAS Blog** – We have posts on a wide range of topics and update this regularly https://ncgas.org/NCGAS_Blog.php#NCGAS%20Blog
- **Look-up resources available through NCGAS –** XSEDE resources - https://www.xsede.org/ecosystem/resources, IU machines- https://kb.iu.edu/d/alde, Science gateways- https://ncgas.org/galaxy.php
- **Request allocation on these machines-** To get access to any of the resources above, here is the allocation request page - https://ncgas.org/hps.php
- **Software available on NCGAS resources-** On the IU/PSC clusters - https://ncgas.org/software-list.php , Jetstream virtual machines- https://ncgas.org/Software%20list%20Jetstream.php
- Email us at help@ncgas.org if you want to be added to the mailing list for updates!