Basics about HGCC

Jingjing Yang

Department of Human Genetics

Login to HGCC

- MAC
  - Use the **Terminal** application
  - Commands: ssh userID@hgcc.genetics.emory.edu
- Windows
  - Use the terminal emulator **PuTTY**
- Outside Emory Network
  - Have **DUO** authorization set up (http://it.emory.edu/security/services/two_factor/)
  - Connect to **VPN** before login (http://it.emory.edu/security/vpn.html)
  - Call Emory IT Support at 404-727-7777 for help with DUO and VPN

Basic Linux Commands

- Show current directory: pwd
- Up-arrow will go to the previous command
- List files/sub-directories under current directory: ls
- Note: all content surrounded by brackets [ ] have to be replaced by your own content, including the brackets
- Find the manual of Linux command, e.g., ls: man [ls]
- Make new directory: mkdir –p [directory]
- Navigate to a directory: cd [directory]
  - Current directory: .
  - Up 1 level directory: ..
  - Home directory: ~
- Rename/Move directory: mv [directory] [destination directory]

Basic Linux Commands

- Define variables in BASH (HGCC): export VARNAME=[value]
  - Use ${VARNAME} as a shortcut for defined value
  - Show variable content: echo ${VARNAME}
- Look at a file
  - Read file in terminal: less, zless, more
  - Press ‘q’ to exit less, zless
  - Print all file content into terminal: cat, zcat
  - zless and zcat are the commands for gzipped files
  - Look at start of a file:
    - Print first K lines: head –n [K]
    - Print all but the last K lines: head –n –[K]
  - Look at end of a file:
    - Print last K lines: tail –n [K]
    - Print all but the first K lines: tail –n +[K]
- Search for pattern(s) in a file
  - Print lines with pattern: grep [pattern] [file]
  - Print lines without pattern: grep –v [pattern] [file]

Basic Linux Commands

- Line count: wc –l [file]
- Sort file: sort
- Stream commands together (vertical line): |
  - Use the output from the command before the vertical line as input for the command after the vertical line
- Extract specific columns: cut –d[Delimiter] –f[field number]

Edit txt files

- Create a new txt file: touch
- Editor tool in terminal: vi, nano
- Basics for vi:
  - https://www.cs.colostate.edu/helpdocs/vi.html
- Basics for nano:
  - https://wiki.gentoo.org/wiki/Nano/Basics_Guide
- Recommend to use Sublime to edit your code on your local computer

Human Genetics Compute Cluster (HGCC)

- HGCC consists of one head node and 9 computation nodes.
- The computation nodes have varying amounts of RAM, CPU (cores), and local scratch space (/scratch)
- Head node is called node00
- Compute nodes are called node01, node02, …
- List all computation nodes: qstat –f
- List other user‘s jobs: qstat –u ‘*‘

Rules about Using HGCC

- **Never run big jobs on head node**
- Submit big jobs to computational nodes (only submit jobs at head node): qsub
- Use command qlogin at head (gateway) node to login to an interactive session
  - Log into specific node: qlogin –l h=[node07.local]
  - Only test your jobs at interactive sessions
- **Use local scratch space**
  - Copy big data to the scratch spaces (/scratch) on each computational node to avoid extensive I/O
  - Remove your data from scratch space when your job is done

Submit Jobs to HGCC

- Submit your job from the Head node (run the job under current working directory (-cwd), with given job name (-N), requesting 1 core (-pe smp 1):
  - qsub –q b.q –cwd –j y –N [jobname] –pe smp 1 [job commands]
  - Note: qsub limit is 500 jobs. Please use array jobs for a large number of jobs (with option –t)
  - See instructions:
    - http://wiki.gridengine.info/wiki/index.php/Simple-Job-Array-Howto
- Check job status: qstat
- Delete a job: qdel [jobid]
- Delete all your jobs: qdel –u [userid]

Key Strategies

- Use shell scripts (see BASH.pptx)
- Commonly used tools are installed as modules (**can only be used after login to an interactive session with command qlogin**)
  - See installed modules/tools : module avail
  - Load a module/tool : module load [software]
  - List loaded modules: module list
  - Unload a module: module unload [software]
  - Unload all loaded modules: module purge
- Do not make another copy of the data on HGCC unless you need to make changes
- Common reference genome data sets on HGCC
  - HGCC shared reference genome data
    - /sw/hgcc/Data

SGE Queues on HGCC

- There are two queues defined on HGCC – b.q and i.q
- b.q
  - For batch (non-interactive) jobs
  - Restricted to node01 – node06
  - Job defaults
    - 1 core / 8GB RAM
    - 240 hours max. run time
  - Requestable resources
    - Cores
    - Run time
  - Memory is not requestable – you get 8 GB / core (See slide on requesting additional resources)
- i.q
  - For interactive jobs, e.g. to run program with a GUI, or requiring command line access
  - Restricted to node07 – node09
  - Job defaults
    - 1 core / 8GB RAM
    - 24 hours max. run time
  - Requestable resources
    - Cores

Requesting additional cores for your job

- To request additional cores
  - `qsub -q b.q -pe smp 4 …`
- Notes
  - Requesting additional cores also provides additional memory
    - 1 core = 8 GB, 2 cores = 16GB, 4 cores = 32GB, …
  - Your program(s) must be able to take advantage of multiple cores or additional memory.
  - You may have to specify this via the program's command line options, e.g. specifying –p option for bowtie2: http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#performance-tuning
  - The smp parallel environment requires that the requested number of cores be free/available on a single node, otherwise you job will not run.
  - Multiple smUsing more cores/memory may not result in a dramatic performance improvement. Think about possibly breaking your analysis into multiple jobs/steps and running those jobs/steps concurrently on multiple nodes. all jobs may be more efficient than a single large job. It also is more user-friendly.

Requesting additional time for your job

- To request additional time
  - `qsub -q b.q -l h_rt=hh:mm:ss ...`
    - hh = hours, mm = minutes, ss = seconds
- Notes:
  - Default run time for batch jobs is 240 hours.
  - This is sufficient for 99.9% of jobs on HGCC. If your job is taking more than 240 hours to run, it's probably stuck and should be terminated.
  - You can also request a shorter run time, e.g. for testing purposes
    - `qsub -q b.q -l h_rt=1:00:00 ...`
    - The above will run your job for one hour then automatically terminate it.