Genome Data Management Workshop

January 22nd-26th 2011

Universidad de Puerto Rico en Mayagüez

GDM Workshop

Day 1

I. DNA sequencing

II. Planning your project

III. Hardware requirements

IV. A primer on Linux

Day 2

V. Genome assembly

Day 3

VI. Genome annotation

Day 4

VII. Perl

VIII. Phylogeny

Day 5

VIII. Phylogeny

WIFI Network code
SquawkinGood

II. Planning your project

3 questions

What is your goal?

What is the best way to achieve it?

How much can you spend?

Evaluate your strengths

Outsourcing

Got DNA/RNA?

Libraries?

Sequencing?

Bioinformatics?

A collaboration?

Additional costs

PCRs

Cloning

Sanger sequencing

Misc. lab costs

III. Hardware requirements

The hidden cost

1 Gigabase of DNA sequence ≈ 1 GB of RAM

32 bits vs 64 bits vs 128 bits

32 bits = 2^{32} = 4 294 967 296 = 4 Gigabytes of RAM

64 bits = 2^{64} = 18 446 744 073 709 551 616 = 16 Exabytes of RAM

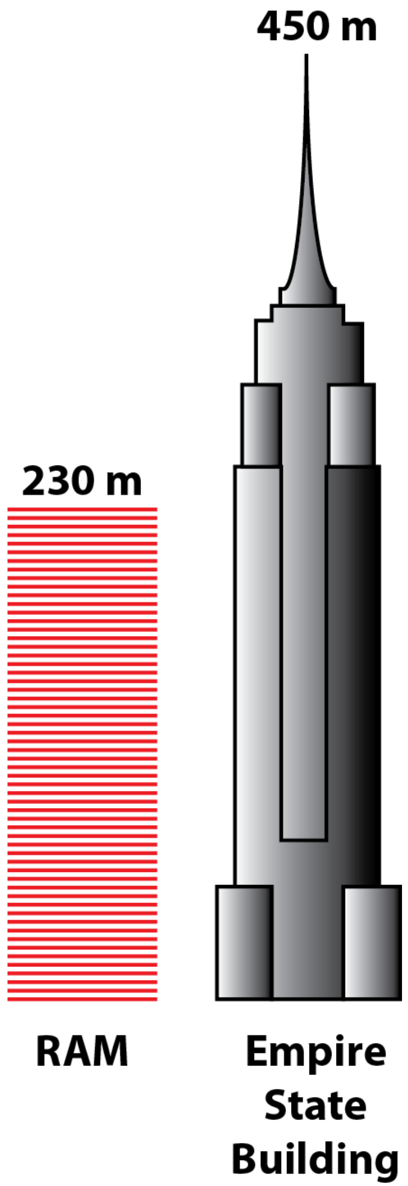
128 bits = 2^{128} = do we even have a number for that? **^ Not quite**

256 TB of RAM

256 TB = 262 144 GB

At current density (8GB per chip) = 32 768 chips

At a thickness of ≈ 7mm, if piled up…



Operating systems (64 bits)

MS Windows 7

Starter	2 GB
Home Basic	8 GB
Home Premium	16 GB
Professional, Enterprise, Ultimate	192 GB

Windows Server 2008 R2

Datacenter, Enterprise	2 TB

Mac OSX

Leopard/Snow Leopard 32/64 bits kernels	unspecified

Linux

44 bits kernel	16 TB
45 bits kernel	32 TB
46 bits kernel	64 TB

Disk space

Data files	1-100 Gigabytes
Temporary files	can easily range in Terabytes
Current SATA drives	3 Terabytes

“Let's face it, we're not changing the world. We're building a product that helps people buy more crap - and watch porn.”

Bill Watkins, Seagate CEO - 2006

I/O speed

Optical media	Not a viable option
SSD	Fast, small, expensive
Hard drive	Slower, bigger, cheaper
RAID*	Software or hardware

*Not the WoW kind…

File compression

Saves on disk space

Faster I/O

Requires more RAM & CPU (at launch)

Works well on text files (*i.e.* most sequencing output files)

The beasts of burden

Supercomputers

Local clusters

Workstations

Mobile workstations

Cloud-based, networked or local?

Computing resources

Compute/Calcul Canada
https://computecanada.org/

TeraGrid
https://www.teragrid.org/

A supercomputer (Le colosse)

CPU	960 nodes Xeon X5560 @ 2.8GHz (7680 cores)
RAM	24GB/node (23TB total)
Disk	500TB Lustre FS

CLUMEQ, Université Laval, Quebec City

Supercomputer/Clusters limitations

RAM per node

Disk quota

Bandwidth

Queuing & CPU time limit

User permissions

IBM BladeCenter HX5 Express MAX5

CPU	2x Intel Xeon X6550 Quad Core 2.0GHz
RAM	**112GB** DDR3 ECC

22 800$ + tax

IBM BladeCenter HX5 Express MAX5

CPU	2x Intel Xeon E7540 Quad Core 2.0GHz
RAM	**64GB** DDR3 ECC

15 000$ + tax

My own rig (custom built)

CPU	2x Intel Xeon E5506 Quad Core 2.13GHz
RAM	**96GB** DDR3 ECC

5 500$ + tax

Backups

DVDs	Too small
Blu-Rays	Might do it
External disks	Decent option
Tape drives	Decent option
NAS	Better option

IV. A primer on Linux

What is Linux?

Free operating system created by Linus Torvalds & Richard Stallman

Similar to UNIX by AT&T (Bell Labs)

Actively developed under the GNU General Public License

Source code is freely available

The Linux distributions

Fedora
http://fedoraproject.org/

Ubuntu
http://www.ubuntu.com/

Red Hat
http://www.redhat.com/

openSUSE
http://www.opensuse.org/

Gentoo
http://www.gentoo.org/

And many more…

How?

- From a live CD/DVD or install CD/DVD
- From within Windows (WUBI)
- From a flash drive
- From the web via HTTP of FTP

The Linux kernel

The command shell

BASH	(Bourne-Again SHell)
TCSH	
DASH	(Debian Almquist SHell)
KSH	

The X interface

The window managers

IceWM
http://www.icewm.org/

Sawfish
http://sawfish.wikia.com/

Fluxbox
http://fluxbox.org/

Enlightenment
http://www.enlightenment.org/

And many more…

The desktop managers

Gnome
http://www.gnome.org/

KDE
http://www.kde.org/

XFCE
http://www.xfce.org/

LXDE
http://lxde.org/

Compilers/interpreters

Source code

Programming languages

The compilers

Binaries

The partitions

Boot

Swap

Root (/)

The user types

Root

Sudoers

Normal users

Groups

The user permissions

The owner	d**rwx**rwxrwx
The group	drwx**rwx**rwx
The others	drwxrwx**rwx**

r = read, w = write, x = execute, - = permission denied

Basic commands

cd	change directory
cp	copy files/directories
mv	move/rename files/directories
ls	list the content of a directory
rm	delete files/directories
*	the wildcard
>	redirect the bash output to a file
>>	append the bash output to a file

Useful commands

pwd	display your current directory
tar	compress files/folders
ln	create aliases (links)
top	show processes
jobs	display current jobs
kill	terminate jobs/processes
screen	detachable Shell
shutdown	stop/reboot computer

Useful tips

Tab-typing	autocomplete your command
cd $HOME	go back to your home directory
cd ~	another way to go back to $HOME
Ctrl+C	abort a process
history	show your last commands
up and down arrows	scroll through previous commands
less/more	reads the 1st lines of a text file
tail	reads the last lines of a text file

Running the analyses

Command lines vs. GUI

Locally

Remotely

SSH & SFTP

Installing additional software

The easy way

The hard way

The Path

A word on dependencies

Need help?

http://embnet.org/en/QuickGuides