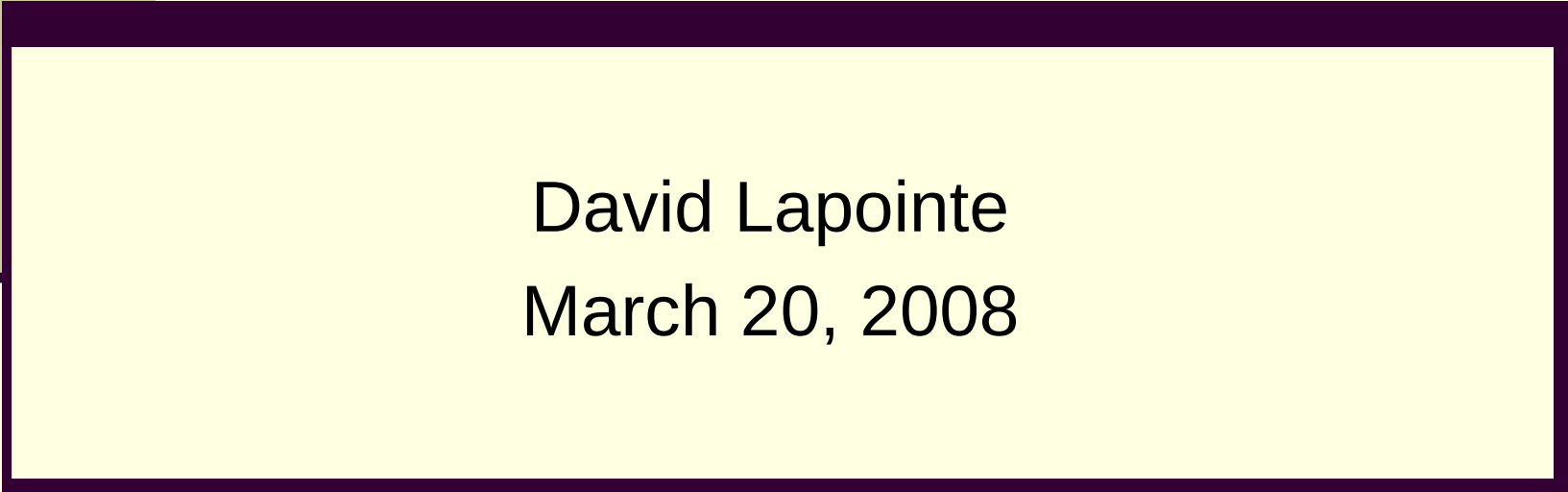




# Solexa Data Crunching



David Lapointe  
March 20, 2008

# Data Crunching

---

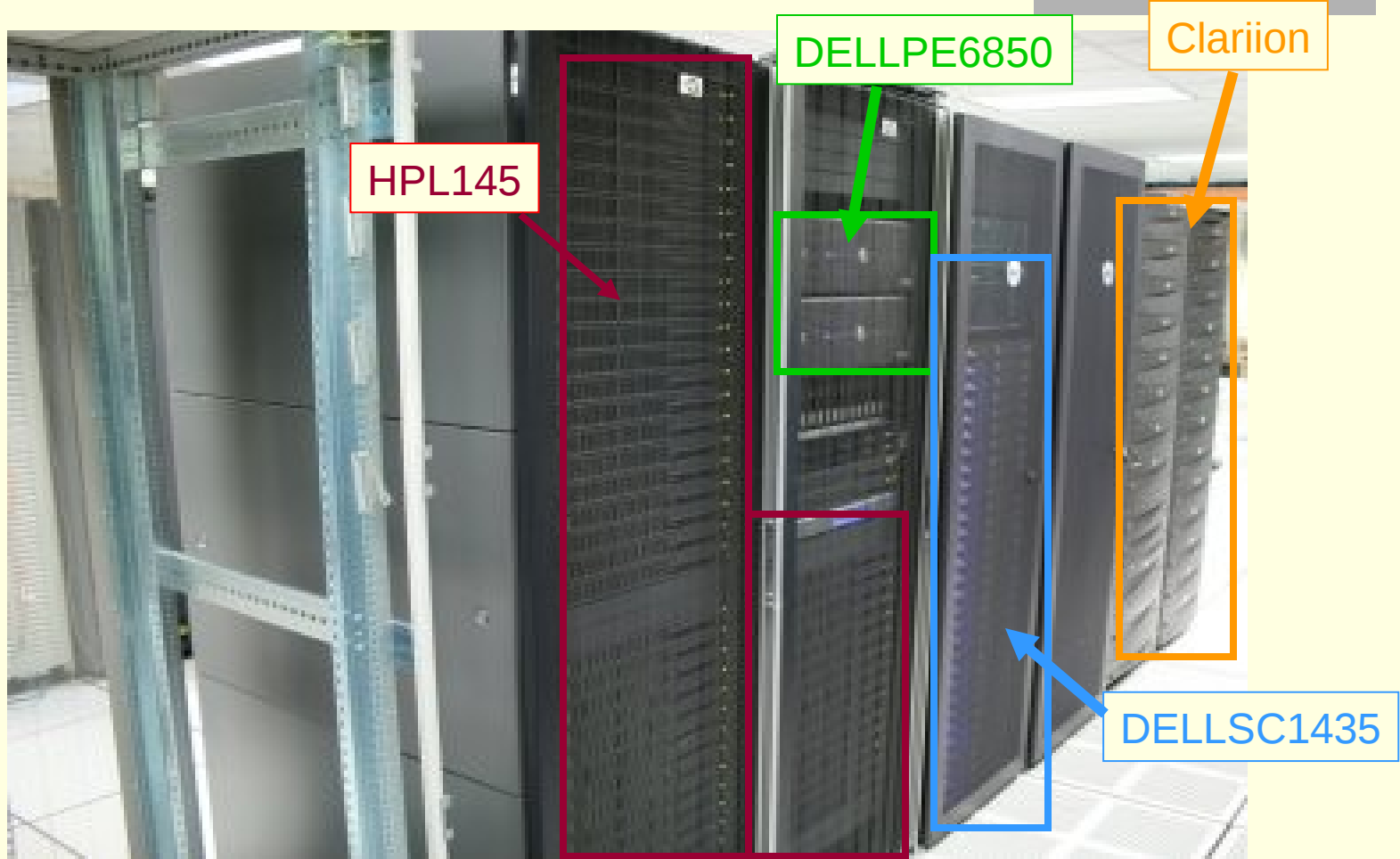
- Binar – A Tour
- Resources on Binar
- How Data is Crunched
- Other Resources

# Binar

---



# Binar: Components



# More Specs

---

- HPL145
  - 2 dual core Opteron/4 Gb mem
  - 55 nodes
- DELLPE6850
  - 4 quad core Intel/64 Gb mem
  - 2 nodes
- DELLSC1435
  - 2 dual core Opteron/4 Gb mem
  - 25 nodes
  - Infiniband capable

# Cluster Use/Etiquette

---

- Use queue system to run jobs
  - Sun Grid Engine queueing
  - Fair share queueing is in use
- File space is not storage
  - Do not store data on the cluster
  - Cluster has 2.0 Tb only for workspace
  - 23 Active Users

# Queues

---

`$qsub script`      basic form

`$qstat`              shows jobs in queue

`$qstat |grep user`    shows jobs for username

Info for Binar and Sun Grid Engine is online at  
<http://inside.umassmed.edu/is/acs/ResearchComputing/researchclusters.aspx>

# Script Template for Queue

---

```
#!/bin/bash
```

```
#$ -S /bin/bash      <- specify shell
```

```
#$ -cwd              <- use current working dir
```

```
#$ ... other sge parameters
```

```
PATH=<path to your program>:$PATH
```

```
export PATH
```

...rest of script ....

Then submit script within working directory

```
$qsub -cwd script
```

# Scripts

---

It is very important to set custom paths, esp for programs that you write, or using bioperl.

`/share/...` is common to all nodes.

`/export/home/<user>` (your home directory) is available to remote nodes.

# Important Directories

---

`/share/apps/bin` commonly used programs

`/share/apps/pipeline/...` Solexa pipeline progs

`/share/nemo/Genomes` contains genome directories formatted for Eland

`/share/nemo/Genomes/hg18`

`/share/nemo/Genomes/mm9`

# Resources on Binar

---

- EMBOSS (similar to GCG, scriptable)
- Clustalw, T\_Coffee - multiple alignment
- Phylip, MrBayes - phylogenetic
- BLAST, mpi-blast
- HMMER
- Bioperl, biopython

These have their own directories.

# Current Genomes on Binar

---

- hg18                    Human (UCSC)
- mm9, mm8            Mouse (UCSC)
- dm5.5, dm5.4        Drosophila (Flybase)
- yeast                    SGD
- z7v                      Zebrafish
- ceWS187                C.Elegans

`/share/nemo/Genomes/xxxx`

# Creating a Genome file for Eland

---

You can run eland against a custom Genome.

2. Each piece ( e.g. a chromosome) must be a separate fasta file.
3. Create a directory for the genome  
`mkdir ~/mygenome`
4. Run squashGenome  
`/share/nemo/pipeline/Eland/squashGenome \  
~/mygenome pathtoFasta/*.fasta`
5. More detail in Pipeline docs

# Pipeline processing

---

- Three phases
  - Image Analysis - Firecrest
  - Base Calling - Bustard
  - Sequence Mapping – Gerald
- Initial Run to generate Cross-talk and offsets
- Run the first two by lane
  - If you specify Genome Gerald can be run
  - If we don't have the genome, let us know what and where, then we can install it.

# Post-processing

---

- When runs are done several files are available
  - quality files, base calls
  - sequences, remapping results
  - summary data
- These are packaged and delivered to you.
- Let me know if you need custom programs for analysis, i.e. ones not delivered with the pipeline

# Resources

---

- Cluster info

<http://inside.umassmed.edu/is/acs/ResearchComputing/researchclusters.aspx>

- Pipeline Info

<http://biotools.umassmed.edu/BIOCORE/pipeline>

- Solexa Google Group

<http://groups.google.com/group/solexa>

- Bioc-Seq (New bioconductor group)

<https://stat.ethz.ch/mailman/listinfo/bioc-sig-sequencing>

- UNIX On-line Help

<http://biotools.umassmed.edu/unixhelp>