Hands-on Bioinformatics on Purdue Supercomputers

Gladys K. Andino, Ph.D.
Senior Scientific Applications Analyst & Bioinformatics Specialist
Founder of Women in HPC at Purdue Research Computing
gandino@purdue.edu

February 12th, 2019
Hands-on

- Data Preprocessing
  - FastQC
  - Trimmomatic
  - Blast
Overview

- Connect to Scholar via
  - MoabXterm, Terminal or Thinlinc
    - Thinlinc: desktop.scholar.rcac.purdue.edu
  - Log in:
    ssh -Y myusername@scholar.rcac.purdue.edu
- Change directory to scratch (cd $RCAC_SCRATCH)
- Check with pwd
- Copy directory class_data from /scratch/scholar/gandino/class_data_data to ./
- Then change the name of the directory to data
- Review and run (.sub files in data )
  - fastqc.sub, trimmomatic.sub, blastx.sub
OVERVIEW

• We will be using the Scholar queue

• [www.rcac.purdue.edu/compute/scholar](http://www.rcac.purdue.edu/compute/scholar)

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Number of Nodes</th>
<th>Cores per Node</th>
<th>Memory per Node</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scholar</td>
<td>8</td>
<td>20</td>
<td>64 GB</td>
</tr>
</tbody>
</table>

• Scholar cluster is open to Purdue classroom instructors from any field whose classes include assignments that could make use of supercomputing.
LOGGING IN – using MobaXterm

WINDOWS

Make sure you start the X server too
LOGGING IN – using MobaXterm

WINDOWS

• Host Name for scholar is “scholar.rcac.purdue.edu”

• Username
LOGGING IN – using a MAC

Mac

• Mac OS X has built in Terminal app that can use SSH

• Open Finder and Go to Applications
LOGGING IN – using a MAC

MAC

- Find Utilities folder, open it, and find Terminal app
LOGGING IN – using a MAC

MAC

• Connect using `ssh -Y username@scholar.rcac.purdue.edu`

For X11 forwarding to work in a Mac make sure you have installed XQuartz tools.
LOGGING IN – using MAC

MAC

• You should now be logged in!
LOGGING IN – using THINLINC

https://www.rcac.purdue.edu/compute/scholar/

https://desktop.scholar.rcac.purdue.edu:300/main/
LOGGING IN – using THINLINC
LOGGING IN – using THINLINC
LOGGING IN – using THINLINC

Welcome to the Rice Cluster

Rice nodes are comprised of dual 10-core Intel Xeon-E5 CPUs with 56 Gbps Infiniband. Your scratch is mounted at the path specified in the RCAC_SCRATCH environment variable.

New mailing list for researchers interested in high performance computing for their work.

info: https://lists.purdue.edu/mailman/listinfo/hpc
subscribe with email to 'hpc-subscribe@lists.purdue.edu'

Scholar Queue Limits are as follows:
Max Jobs in Queue per User = 5
Max Walltime per Job = 40 Hours

Latest module updates:
gcc 6.3.0 was updated on Wed Aug 2 09:41
gcc 7.1.0 was updated on Wed Aug 2 10:11
gcc 6.3.0 was updated on Wed Aug 2 10:11
qt 5.8.0 was updated on Thu Aug 3 19:11

(scholar-fe83)
Sample Preparation
Illumina Sequencing
Demultiplexing

Data Cleaning

Raw FastQ

Cleaned FastQ

FastQC

Tophat

Trinity

Reference Genome (fasta)

Reference Annotation (GTF)

Cufflinks

Merged Annotation (GTF)

Cuffdiff

Normalized counts DEGs

Cummerbund

HTSeq

Raw Counts

DeSeq2

DEGs

Trinity

RSEM

Pathway analysis
GO analysis

Data Preprocessing

Sample Preparation
Illumina Sequencing
Demultiplexing

Workflow
### Sequencing Basics

**Fastq format**

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>@HISEQ02:319: C22FKACXX: 2: 1101: 1699: 1972 1:N:0: GTAGAG</td>
<td></td>
</tr>
<tr>
<td>GACCCATCCATTGTTGGACAGCTGAAGACGGGACGATCGTGCTCGTGGTTTTGAATGCGAGAATCCCTGCAGAGGCTGCCTGCTTCGGNNNNNNNNNNNTCCCTGACAGCC+</td>
<td></td>
</tr>
<tr>
<td>CCCCCHHHHHJIJJJGIIJJJJJJJJIIJJIIIIAIJJEHHHHFFFDCCDDDDDDCDDDDDDDDDBBBDDDDDCDDDB</td>
<td></td>
</tr>
</tbody>
</table>

#### ASCII Basics

- `I` = ASCII 73
- `#` = ASCII 35
- `Quality = 73 - 33 = 40`
- `Quality = -10 \log_{10} \epsilon`
- `\epsilon = 10^{-4}`

#### Quality Calculation

- `Q = 35 - 33 = 2`
- `\epsilon = 10^{-0.2} = 0.63` totally bogus

### ASCII BASE=33 Illumina, Ion Torrent, PacBio and Sanger

<table>
<thead>
<tr>
<th>Q</th>
<th>P_error</th>
<th>ASCII</th>
<th>Q</th>
<th>P_error</th>
<th>ASCII</th>
<th>Q</th>
<th>P_error</th>
<th>ASCII</th>
<th>Q</th>
<th>P_error</th>
<th>ASCII</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.00000</td>
<td>33 !</td>
<td>11</td>
<td>0.07943</td>
<td>44 ,</td>
<td>22</td>
<td>0.00631</td>
<td>55 7</td>
<td>33</td>
<td>0.00050</td>
<td>66 B</td>
</tr>
<tr>
<td>1</td>
<td>0.79433</td>
<td>34 &quot;</td>
<td>12</td>
<td>0.06310</td>
<td>45 -</td>
<td>23</td>
<td>0.00501</td>
<td>56 8</td>
<td>34</td>
<td>0.00040</td>
<td>67 C</td>
</tr>
<tr>
<td>2</td>
<td>0.63096</td>
<td>35 #</td>
<td>13</td>
<td>0.05012</td>
<td>46 .</td>
<td>24</td>
<td>0.00398</td>
<td>57 9</td>
<td>35</td>
<td>0.00032</td>
<td>68 D</td>
</tr>
<tr>
<td>3</td>
<td>0.50119</td>
<td>36 $</td>
<td>14</td>
<td>0.03981</td>
<td>47 /</td>
<td>25</td>
<td>0.00316</td>
<td>58 :</td>
<td>36</td>
<td>0.00025</td>
<td>69 E</td>
</tr>
<tr>
<td>4</td>
<td>0.39811</td>
<td>37 %</td>
<td>15</td>
<td>0.03162</td>
<td>48 0</td>
<td>26</td>
<td>0.00251</td>
<td>59 ;</td>
<td>37</td>
<td>0.00020</td>
<td>70 F</td>
</tr>
<tr>
<td>5</td>
<td>0.31623</td>
<td>38 &amp;</td>
<td>16</td>
<td>0.02512</td>
<td>49 1</td>
<td>27</td>
<td>0.00200</td>
<td>60 &lt;</td>
<td>38</td>
<td>0.00016</td>
<td>71 G</td>
</tr>
<tr>
<td>6</td>
<td>0.25119</td>
<td>39 '</td>
<td>17</td>
<td>0.01995</td>
<td>50 2</td>
<td>28</td>
<td>0.00158</td>
<td>61 =</td>
<td>39</td>
<td>0.00013</td>
<td>72 H</td>
</tr>
<tr>
<td>7</td>
<td>0.19953</td>
<td>40 (</td>
<td>18</td>
<td>0.01585</td>
<td>51 3</td>
<td>29</td>
<td>0.00126</td>
<td>62 &gt;</td>
<td>40</td>
<td>0.00010</td>
<td>73 I</td>
</tr>
<tr>
<td>8</td>
<td>0.15849</td>
<td>41 )</td>
<td>19</td>
<td>0.01259</td>
<td>52 4</td>
<td>30</td>
<td>0.00100</td>
<td>63 ?</td>
<td>41</td>
<td>0.00008</td>
<td>74 J</td>
</tr>
<tr>
<td>9</td>
<td>0.12589</td>
<td>42 *</td>
<td>20</td>
<td>0.01000</td>
<td>53 5</td>
<td>31</td>
<td>0.00079</td>
<td>64 @</td>
<td>42</td>
<td>0.00006</td>
<td>75 K</td>
</tr>
<tr>
<td>10</td>
<td>0.10000</td>
<td>43 +</td>
<td>21</td>
<td>0.00794</td>
<td>54 6</td>
<td>32</td>
<td>0.00063</td>
<td>65 A</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data Preprocessing

Adapter/quality trimming

Before

After
Data Preprocessing

Cleaning

- Technical
  - Adapters
  - Read quality
  - PCR artifacts

- Contaminants
  - mitochondrial/chloroplast
  - ribosomal RNA
  - phiX174
  - environmental
  - other?
#!/bin/sh -l
#PBS -N fastqc
#PBS -q scholar
#PBS -l nodes=1:ppn=2
#PBS -l walltime=00:10:00
#PBS -l naccesspolicy=shared

module purge
module load bioinfo
module load fastqc
module list

cd $PBS_O_WORKDIR
pwd
cat fastqc.sub
date +"%d %B %Y %H:%M:%S"
echo " "
# fastq reads mites
fastqc mites_R1.fastq \ 
mites_R2.fastq \ 
-t 2 -o ./
echo " "
date +"%d %B %Y %H:%M:%S"
FastQC: Results

Note: the Job ID number
fastqc.o21505
fastqc.e21505
FastQC: Results

```
gandino@scholar-fe03:/scratch/scholar/gandino/data $ qstat -u gandino

scholar-adm.rcac.purdue.edu:

<table>
<thead>
<tr>
<th>Job ID</th>
<th>Username</th>
<th>Queue</th>
<th>Jobname</th>
<th>SessID</th>
<th>NDS</th>
<th>TSK</th>
<th>Memory</th>
<th>Req'd Time</th>
<th>S</th>
<th>Req'd Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>21505.scholar-adm.rcac</td>
<td>gandino</td>
<td>scholar</td>
<td>fastqc</td>
<td>5177</td>
<td>1</td>
<td>2</td>
<td>--</td>
<td>00:10:00</td>
<td>R</td>
<td>00:00:13</td>
</tr>
</tbody>
</table>
```

Note: the Job ID number will be assigned to the o & e files

fastqc.o21505
fastqc.e21505
FastQC: Results

- Enabling GUI interface

```
$ module load bioinfo fastqc
$ fastqc
```
Trimmomatic - PE

- Trimmomatic: A flexible read trimming tool for Illumina NGS data
  
  http://www.usadellab.org/cms/?page=trimmomatic

- Paired End Mode:
- Single End Mode:

Usage:

Trimmomatic PE [-threads <threads>] [-phred33 | -phred64] [-trimlog <logFile>] <input 1> <input 2> <paired output 1> <unpaired output 1> <paired output 2> <unpaired output 2> <step 1> ...

Trimmomatic - PE

- **ILLUMINACLIP**: Cut adapter and other illumina-specific sequences from the read.
- **SLIDINGWINDOW**: Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold.
- **LEADING**: Cut bases off the start of a read, if below a threshold quality
- **TRAILING**: Cut bases off the end of a read, if below a threshold quality
- **MINLEN**: Drop the read if it is below a specified length
- **TOPHRED33**: Convert quality scores to Phred-33

This will perform the following:

- Remove adapters (**ILLUMINACLIP**:illumina-adap.fa:2:35:15)
- Remove leading low quality or N bases (below quality 7) (**LEADING**:7)
- Remove trailing low quality or N bases (below quality 7) (**TRAILING**:7)
- Scan the read with a 4-base wide sliding window, cutting when the average quality per base drops below 15 (**SLIDINGWINDOW**:4:15)
- Drop reads below the 30 bases long (**MINLEN**:30)
#!/bin/sh -l
#PBS -N trimmomatic
#PBS -q scholar
#PBS -l nodes=1:ppn=8
#PBS -l walltime=00:15:00
#PBS -l naccesspolicy=shared

module purge
module load bioinfo
module load trimmomatic
module list

cd $PBS_O_WORKDIR
pwd

# Example PE end mode
trimmomatic PE -threads 8 \
-phred33 \
-trimlog trimmo_mites-reads.log \
mites_R1.fastq \
mites_R2.fastq \
output_mites_R1_paired.fq \
output_mites_R1_unpaired.fq \
output_mites_R2_paired.fq \
output_mites_R2_unpaired.fq \
ILLUMINACLIP:illumina-adap.fa:2:35:15 \
LEADING:7 \
TRAILING:7 \
SLIDINGWINDOW:4:15 \
MINLEN:30
Run fastQC: using the output files from trimmomatic

$ module load bioinfo
$ fastqc

Use > open to select the sequence file you want to check
- Select output_mites_R1_paired.fq
**BLAST: blastx**

- Basic Local Alignment Search Tool
- Blastx: search protein databases using a translated nucleotide query
BLAST: blastx.sub

- Connect to Scholar
- Run blastx (qsub jobfile)
- Briefly discuss results

```
#!/bin/sh -l
#PBS -N blastx
#PBS -q scholar
#PBS -l nodes=1:ppn=4
#PBS -l walltime=00:15:00
#PBS -l naccesspolicy=shared

module purge
module load bioinfo
module load blast
module list

cd $PBS_O_WORKDIR
pwd

cat blastx.sub
date +"%d %B %Y %H:%M:%S"
echo " "

# blastx command example
blastx -query mites_seq.fasta \
-db swissprot -num_threads 4 \
-out mites_seq_out.fmt6 \
-outfmt 6 \
-evalue 1E-06 \
-max_target_seqs 5

echo " "
date +"%d %B %Y %H:%M:%S"
```
### BLAST: results

- *-outfmt 6 (tabular format)*

<table>
<thead>
<tr>
<th>qseqid</th>
<th>sseqid</th>
<th>%ident</th>
<th>len</th>
<th>mismat</th>
<th>gapopen</th>
<th>qstart</th>
<th>qend</th>
<th>sstart</th>
<th>send</th>
<th>Eval</th>
<th>qlen</th>
</tr>
</thead>
<tbody>
<tr>
<td>vjasml_127944</td>
<td>Q9XSC3.1</td>
<td>59.765</td>
<td>425</td>
<td>136</td>
<td>7</td>
<td>203</td>
<td>1417</td>
<td>495</td>
<td>904</td>
<td>8.04E-166</td>
<td>523</td>
</tr>
<tr>
<td>vjasml_127944</td>
<td>Q498F0.1</td>
<td>60.563</td>
<td>426</td>
<td>138</td>
<td>8</td>
<td>203</td>
<td>1432</td>
<td>498</td>
<td>909</td>
<td>2.65E-165</td>
<td>522</td>
</tr>
<tr>
<td>vjasml_127944</td>
<td>Q5JSH3.1</td>
<td>59.390</td>
<td>426</td>
<td>146</td>
<td>6</td>
<td>203</td>
<td>1432</td>
<td>496</td>
<td>910</td>
<td>7.67E-164</td>
<td>518</td>
</tr>
<tr>
<td>vjasml_127944</td>
<td>Q6NVE8.1</td>
<td>60.808</td>
<td>421</td>
<td>138</td>
<td>7</td>
<td>203</td>
<td>1417</td>
<td>498</td>
<td>907</td>
<td>2.26E-163</td>
<td>517</td>
</tr>
<tr>
<td>vjasml_127944</td>
<td>Q9R037.1</td>
<td>60.817</td>
<td>416</td>
<td>136</td>
<td>7</td>
<td>203</td>
<td>1402</td>
<td>491</td>
<td>895</td>
<td>3.76E-161</td>
<td>511</td>
</tr>
</tbody>
</table>
Questions?