Teaching Computational Genomics:
A Tale of Tables and Tests

Texas A&M Research Computing Symposium
2021-05-24

Rodolfo Aramayo - Department of Biology - Texas A&M University
The Genomic Revolution

Landmarks in genetics and genomics

- **1990**: The Human Genome Project (HGP) launched in the United States
- **1991**: First US genome mapping project
- **1993**: First generation human genetic map developed
- **1994**: The HGP's human genetic mapping goal achieved
- **1995**: First human genome project established
- **1997**: US National Center for Human Genome Research becomes the National Human Genome Research Institute (NHGRI)
- **1998**: Informed consent model of 35,000 genes enters into human genome project
- **2001**: Sequence of first human chromosome-enzymes completed
- **2003**: Finished version of human genome sequence completed

**Human Genome Project**

**Draft version of human genome sequence published**

- **2000**: Draft version of human genome sequence completed and published
- **2002**: Draft version of human genome sequence completed and published
- **2003**: Finished version of human genome sequence completed

**First bacterial genome (Haemophilus influenzae)** sequenced

**Human Genome Project**

- **1990**: The Human Genome Project (HGP) launched in the United States
- **1991**: First US genome mapping project
- **1993**: First generation human genetic map developed
- **1994**: The HGP's human genetic mapping goal achieved
- **1995**: First human genome project established
- **1997**: US National Center for Human Genome Research becomes the National Human Genome Research Institute (NHGRI)
- **1998**: Informed consent model of 35,000 genes enters into human genome project
- **2001**: Sequence of first human chromosome-enzymes completed
- **2003**: Finished version of human genome sequence completed

**Human Genome Project**

- **Draft version of human genome sequence published**
- **2000**: Draft version of human genome sequence completed and published
- **2002**: Draft version of human genome sequence completed and published
- **2003**: Finished version of human genome sequence completed

**First bacterial genome (Haemophilus influenzae)** sequenced

**Human Genome Project**

- **1990**: The Human Genome Project (HGP) launched in the United States
- **1991**: First US genome mapping project
- **1993**: First generation human genetic map developed
- **1994**: The HGP's human genetic mapping goal achieved
- **1995**: First human genome project established
- **1997**: US National Center for Human Genome Research becomes the National Human Genome Research Institute (NHGRI)
- **1998**: Informed consent model of 35,000 genes enters into human genome project
- **2001**: Sequence of first human chromosome-enzymes completed
- **2003**: Finished version of human genome sequence completed

**Human Genome Project**

- **Draft version of human genome sequence published**
- **2000**: Draft version of human genome sequence completed and published
- **2002**: Draft version of human genome sequence completed and published
- **2003**: Finished version of human genome sequence completed

**First bacterial genome (Haemophilus influenzae)** sequenced
Milestones In Genomic Sequencing

We Sequenced our own Genome!

**Milestone 1  2001**

**The Human Genome Project**

Launched in 1990, the Human Genome Project set out to identify the order, that is, sequence, of all DNA bases to obtain the ‘genetic blueprint’ of humans. In 2001, two pivotal publications reported the first draft of the human genome, obtained by shotgun sequencing, setting the stage for the genomic era. The second phase of the project, which moved from the draft to an essentially finished reference genome, was completed in 2003. [Read more.]

By Caroline Barranco

https://www.nature.com/immersive/d42859-020-00099-0/index.html
Milestones In Genomic Sequencing

We sequenced what was previously impossible!

**Milestone 2  2004**

**Sequencing the unculturable majority**

Two key studies unlocked the field of metagenomics – the reconstruction of microbial communities from sequencing data – by providing approaches for unbiased, culture-independent analysis of DNA directly from environmental samples using sequencing technologies. [Read more.]

By Iain Dickson

Credit: Science Photo Library / Alamy Stock Photo

https://www.nature.com/immersive/d42859-020-00099-0/index.html
We invented new sequencing technologies!

**Milestone 3 2005**

**Sequencing – the next generation**

Two revolutionary studies introduced high-throughput, massively parallel sequencing technologies able to sequence a bacterial genome at a fraction of the cost and time of traditional Sanger sequencing techniques. [Read more.](https://www.nature.com/immersive/d42859-020-00099-0/index.html)

By Joseph Willson

Credit: Zoran Obradovic / Alamy Stock Photo
We understood the need and value of being inclusive!

Milestone 5  2008

The dawn of personal genomes

Two studies reported the genomes of an African individual and an Asian individual, respectively, using a new massively parallel sequencing approach based on reversible terminator dyes. Demonstrating the feasibility and resource value of human genome sequences, these studies and the technology they presented paved the way for population-scale genome sequencing.

Read more.

By Darren Burgess

https://www.nature.com/immersive/d42859-020-00099-0/index.html
Milestones In Genomic Sequencing

Sequencing changed our understanding of disease

Milestone 6  2008

A sequencing revolution in cancer

Ley et al. presented the first whole-genome sequence of a cytogenetically normal acute myeloid leukaemia sample, showing that cancer genome sequencing can identify disease-associated mutations and druggable targets, offering promise for personalized medicine approaches.

Read more.

By Safia Danovi

Credit: Sabena Jane Blackbird / Alamy Stock Photo

https://www.nature.com/immersive/d42859-020-00099-0/index.html
Milestones In Genomic Sequencing

Sequencing changed our understanding of Gene expression and Gene models

Milestone 7  2008

Transcriptomes – a new layer of complexity

A series of milestone publications reported the development of high-throughput sequencing of whole transcriptomes, known as RNA sequencing (RNA-seq), across different species. Read more.

By Margot Brandt

Credit: Y H Lim / Alamy Stock Photo

https://www.nature.com/immersive/d42859-020-00099-0/index.html
We invented even better sequencing technologies!

Milestone 8 2009
Long reads become a reality

Long-read sequencing technologies began to shed light on hidden parts of the human genome by sealing gaps in existing assemblies, allowing modified bases to be detected on native DNA or RNA, and revealing the complexity of the transcriptome. Read more.

By Ivanka Kamenova

Credit: Zoonar GmbH / Alamy Stock Photo
Milestones In Genomic Sequencing

We developed single cell Genomics!

**Milestone 11 2009**

**Sequencing one cell at a time**

Moving from genomic analysis of tissues or cells in bulk to performing single-cell sequencing provided a whole new perspective on gene regulation, cell-to-cell heterogeneity and developmental or disease processes. The difficulty of performing analyses at such resolution required many experimental and computational innovations. [Read more.](https://www.nature.com/immersive/d42859-020-00099-0/index.html)

By Aline Lückgen

Credit: Greenshoots Communications / Alamy Stock Photo
Milestones In Genomic Sequencing

We started sequencing the genomes of our ancestors!

Milestone 12 2010

Waking the dead: sequencing archaic hominin genomes

The publication of the first draft genome of a Neanderthal in 2010 marked a turning point for the palaeogenomics field, making it possible to assemble an ancient genome from next-generation sequencing reads by overcoming previous limitations in ancient DNA research such as limited starting material, contamination and degradation. Read more.

By Rebecca Furlong

https://www.nature.com/immersive/d42859-020-00099-0/index.html
We re-defined the meaning of reference genomes!

**Milestone 15  2014**

**Pan-genomes: moving beyond the reference**

Pan-genome studies in a variety of species — from microorganisms to plants to humans — have shown that a large amount of genetic variation can be found in the dispensable genome. This observation has called into question our reliance on single reference genomes for assembling and analysing genomes. Read more.

By Dominique Morneau

Credit: Art of Food / Alamy Stock Photo

https://www.nature.com/immersive/d42859-020-00099-0/index.html
Milestones In Genomic Sequencing

We learned how to sequence a chromosome with no gaps!

Milestone 17 2020

Filling in the gaps telomere to telomere

2020 saw the publication of the first gapless, telomere-to-telomere assembly of a human chromosome, the X chromosome. This discovery brought together sequencing technologies and computational tools that had been developed in the preceding decade. Read more.

By Katherine Wrighton

Credit: PORNCHAI SODA / Alamy Stock Photo

https://www.nature.com/immersive/d42859-020-00099-0/index.html
The Result was...

2009

An explosion of computational tools

As genome sequencing became more affordable and widespread, its applications rapidly expanded, driving the development of new computational tools to accommodate the requirements of transcriptomics, metagenomics or genetic variant discovery. Read mapping tools such as Bowtie and BWA or the splice-aware aligner TopHat were able to align millions of short reads to the reference genome, and downstream analysis software, such as SAMtools and BreakDancer, facilitated the detection of genetic variants.

Related article: Repetitive DNA and next-generation sequencing: computational challenges and solutions

https://www.nature.com/immersive/d42859-020-00099-0/index.html
The Price of Genomic Sequencing is Dropping...
Genomic Sequencing Is Generating a Problem

The Volume of Incoming Data is Overwhelming
The Problem

Databases are Growing at an Amazing Rate
The Problem

Databases are Growing at an Amazing Rate

*Homo sapiens*

<table>
<thead>
<tr>
<th>Database</th>
<th>Access</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>public</td>
<td>controlled</td>
</tr>
<tr>
<td>BioSample</td>
<td>3,091,653</td>
<td>651,138</td>
</tr>
<tr>
<td>BioProject</td>
<td>32,639</td>
<td>825</td>
</tr>
<tr>
<td>dbGaP</td>
<td></td>
<td>18</td>
</tr>
<tr>
<td>GEO Datasets</td>
<td>806,746</td>
<td></td>
</tr>
</tbody>
</table>

As of: 2021-05-18

The Problem

The Reproducibility Versus Replicability Crisis

Reproducibility of Scientific Results

First published Mon Dec 3, 2018

The terms “reproducibility crisis” and “replication crisis” gained currency in conversation and in print over the last decade (e.g., Pashler & Wagenmakers 2012), as disappointing results emerged from large scale reproducibility projects in various medical, life and behavioural sciences (e.g., Open Science Collaboration, OSC 2015). In 2016, a poll conducted by the journal Nature reported that more than half (52%) of scientists surveyed believed science was facing a “replication crisis” (Baker 2016). More recently, some authors have moved to more positive terms for describing this episode in science; for example, Vazire (2018) refers instead to a “credibility revolution” highlighting the improved methods and open science practices it has motivated.

The crisis often refers collectively to at least the following things:

a. the virtual absence of replication studies in the published literature in many scientific fields (e.g., Makel, Plucker, & Hegarty 2012),

b. widespread failure to reproduce results of published studies in large systematic replication projects (e.g., OSC 2015; Begley & Ellis 2012),

c. evidence of publication bias (Fanelli 2010a),

d. a high prevalence of “questionable research practices”, which inflate the rate of false positives in the literature (Simmons, Nelson, & Simonsohn 2011; John, Loewenstein, & Prelec 2012; Agnoli et al. 2017; Fraser et al. 2018), and

e. the documented lack of transparency and completeness in the reporting of methods, data and analysis in scientific publication (Bakker & Wicherts 2011; Nuijten et al. 2016).

https://plato.stanford.edu/entries/scientific-reproducibility/
The Problem

There is a Divide Between Biology, Computational Sciences and Statistics
And
Biology is developing at an amazing rate, becoming increasingly interdisciplinary

Question

How do you train students for the future, knowing what we know about the past?

One Answer…

You teach them to Think!…Inter-disciplinarily
How do we do that?

Where do we start?
A Solution...

To Gap The Divide Between Biology, Computer Sciences and Statistics

We should start by Teaching Hypothesis-driven Computational Genomics Courses
A Solution…

• Computational Genomics gaps the divide between Biology, Statistics and Computational Sciences

• Computational experiments can and should be designed like any other wet-lab experiment

• The belief that data processing is separated, or different from, wet-lab experiments is biased, damaging and wrong

• The rapid evolution and availability of new algorithms, and computational tools make previous data analyses potentially obsolete

• This generates the need to re-test and re-analyze already produced genome data, and provide us with a unique opportunity for training
A Solution...

BIOL350
COMPUTATIONAL GENOMICS

BIOL650
GENOMICS

BIOL647
DIGITAL BIOLOGY

Undergraduate Level
Started 2010

Graduate Level
Started 2000

Graduate Level
Started 2012
About Computational Genomics (BIOL350) and Genomics (BIOL650)

- **Outside the Firewall**
  - Students
  - Internet
  - GitHub (Class Notes and Class-related Issues)
  - GDrive (Exams and Final Reports)

- **Inside the Firewall**
  - SuperComputer
  - Galaxy (Class Work)

- **VPN**
About Computational Genomics (BIOL350) and Genomics (BIOL650)

Web-Based Computational Tools

Reveille
Deployed 2016
Ada

Galaxy

Kaiser
Deployed 2020
Terra
History of Genomics

Construct clone map and select mapped clones

Generate several thousand sequence reads per clone

Assemble

Generate tens of millions of sequence reads

Assemble

https://www.nature.com/articles/35084503.pdf
About Computational Genomics (BIOL350) and Genomics (BIOL650)

Library Construction → Sequencing → Assembly

- Sequence reads:
  CAGCTGTCCTCAGATGAC AACCTTCCTCCCAGCT TCCGCTTCAGCTCAAGACCTAATCTTC
  GGCTTTTGGGCTC TCCGCTTCAGCTCAAGACCTAATCTTC TCCGCTTCAGCTCAAGACCTAATCTTC
  CAGATGACGCC CCGGCTTTTGGGCTCCGCTTCAGCTCAAGACCTAATCTTC GGGCTCCGCTTCAGCTCAAGACCTAATCTTC
  GGCTTTTGGGCTC CAGATGACGCC

- Match up overlaps:
  CCGGCTTTTGGGCTC CCGGCTTTTGGGCTC CCGGCTTTTGGGCTC CCGGCTTTTGGGCTC
  TCCGCTTCAGCTCAAGACCTAATCTTC TCCGCTTCAGCTCAAGACCTAATCTTC TCCGCTTCAGCTCAAGACCTAATCTTC
  CCGGCTTTTGGGCTC CAGATGACGCC

- Contig:
  CCGGCTTTTGGGCTC CCGGCTTTTGGGCTC CCGGCTTTTGGGCTC CCGGCTTTTGGGCTC
  CAGATGACGCC

http://training.ensembl.org/events/2021/2021-05-18-OpenVirtualBrowser_May
About Computational Genomics (BIOL350) and Genomics (BIOL650)

Genome Annotation

Gene Models

Gene View

Non-coding exon

Non-coding gene transcript

Non-coding transcript of a coding gene

Intron

Coding exon

Merged transcripts

Protein coding transcript

http://training.ensembl.org/events/2021/2021-05-18-OpenVirtualBrowser_May
About Computational Genomics (BIOL350) and Genomics (BIOL650)

Databases

How do I use BioMart?

The four steps

1. Choose database and species
2. Narrow down the dataset
3. Specify your output
4. View and export

http://training.ensembl.org/events/2021/2021-05-18-OpenVirtualBrowser_May
Patterns as Formulas

Patterns

Phe or Tyr
Cys
not Val or Ala
three His

\[ [FY] - x - C - x(2) - \{VA\} - x - H(3) \]

any amino acid
any two amino acids
any amino acid

About Computational Genomics (BIOL350) and Genomics (BIOL650)

Computational Arithmetics

Only records that are joined (inner join):

Overlapping Intervals:

Intervals with no overlap:

Non-overlapping pieces of intervals:

Concatenated intervals

Dataset

Merged intervals
About Computational Genomics (BIOL350) and Genomics (BIOL650)

Comparative Genomics

By Taxa

- Ostelemur gambertii
- Microcebus murinus
- Propithecus coquerelii
- Propithecus verreauxi
- Callicebus jacchus
- Saimiri boliviensis boliviensis
- Cebus imitator
- Aotus nancymaeae
- Cercopithecus alyx
- Mandrillus leucophaeus
- Papio anubus
- Macaca fascicularis
- Macaca mulatta
- Macaca nemestrina
- Cercopithecus sabaeus
- Rhinopithecus bieti
- Rhinopithecus roxellana
- Pongo abelii
- Gorilla gorilla gorilla
- Pan paniscus
- Pan troglodytes
- Homo sapiens
- Nomascus leucogenys
- Cebus pygerythrus

Shared Syntheny

Conserved order of aligned homologous genomic blocks between species (irrespective of orientation):

Homology Relationships

- Orthologues
  - Genes emerged through a speciation event, e.g., C1 and H1; C2 and M; H2 and M
  - 1-to-1: C1 and H1
  - 1-to-many: M and H1, H2
- Paralogues
  - Genes emerged through a duplication event, e.g., C1 and C2, H1 and H2

Gene Gain/Loss Tree

http://training.ensembl.org/events/2021/2021-05-18-OpenVirtualBrowser_May
Final Project Hypothesis:
The number of conserved proteins is directly proportional to the evolutionary distance between the proteomes tested

The Final Project workhorse is the Reverse-Blast-Hit (RBH) Algorithm

All students get the same (control) proteome and then each student gets two different unique proteomes
About Computational Genomics (BIOL350) and Genomics (BIOL650)

Key Concepts

Intersections AND/OR/NOT

Datasets Combinations

Gene Polarity

Gene Models

Proteome A Proteome B Proteome C

Proteome A Proteome B

Proteome A Proteome B Proteome C

Proteome A Proteome B

Proteome A Proteome B

Proteome A Proteome B

Proteome A Proteome B

Databases Relationships

One to One

One to Many

Many to Many
Key Concepts

Relationships: Tables to Graphs

The Canonical Gene

Table containing 12 fields (columns) and 15 records (rows or lines)

Field 01
Record 01

Sorting by Colors
On Field 04

Table containing 12 fields (columns) and 15 records (rows or lines)

Field 01
Record 08

New Head -2

New Tail -4
About Computational Genomics (BIOL350) and Genomics (BIOL650)

Google Drive
Or GITHUB

<table>
<thead>
<tr>
<th>Google Drive</th>
<th>Or</th>
<th>GITHUB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Download Question(s)</td>
<td>Upload Answers(s)</td>
<td></td>
</tr>
</tbody>
</table>

Databases

<table>
<thead>
<tr>
<th>Databases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Send Queries</td>
</tr>
<tr>
<td>Send</td>
</tr>
</tbody>
</table>

SuperComputer Galaxy

<table>
<thead>
<tr>
<th>SuperComputer Galaxy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Send</td>
</tr>
</tbody>
</table>

Process Data

<table>
<thead>
<tr>
<th>Process Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Send</td>
</tr>
</tbody>
</table>

Grading

<table>
<thead>
<tr>
<th>Grading</th>
</tr>
</thead>
<tbody>
<tr>
<td>History</td>
</tr>
<tr>
<td>search datasets</td>
</tr>
</tbody>
</table>

Exam01_Assemblies
35 shown, 3 hidden
83.81 MB

<table>
<thead>
<tr>
<th>Exam01_Assemblies</th>
</tr>
</thead>
<tbody>
<tr>
<td>38: _____________</td>
</tr>
<tr>
<td>37: Sort on data 34</td>
</tr>
<tr>
<td>36: Sort on data 33</td>
</tr>
<tr>
<td>35: Sort on data 32</td>
</tr>
<tr>
<td>34: Compute sequence length on data 28</td>
</tr>
<tr>
<td>33: Compute sequence length on data 20</td>
</tr>
<tr>
<td>32: Compute sequence length on data 12</td>
</tr>
<tr>
<td>31: _____________</td>
</tr>
<tr>
<td>29: VelvetOptimiser on data 5 and data 6: Contig Stats</td>
</tr>
<tr>
<td>28: VelvetOptimiser on data 5 and data 6: Contigs</td>
</tr>
<tr>
<td>27: FastQC on data 6: RawData</td>
</tr>
<tr>
<td>26: FastQC on data 6: Webpage</td>
</tr>
</tbody>
</table>

Text Messaging

<table>
<thead>
<tr>
<th>Text Messaging</th>
</tr>
</thead>
<tbody>
<tr>
<td>Send</td>
</tr>
<tr>
<td>Receive</td>
</tr>
</tbody>
</table>

Individual Processes

<table>
<thead>
<tr>
<th>Individual Processes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student A</td>
</tr>
<tr>
<td>Receive</td>
</tr>
<tr>
<td>Send</td>
</tr>
<tr>
<td>Handle Quesitons</td>
</tr>
<tr>
<td>Receive</td>
</tr>
<tr>
<td>Send</td>
</tr>
<tr>
<td>Student B</td>
</tr>
<tr>
<td>Receive</td>
</tr>
<tr>
<td>Send</td>
</tr>
<tr>
<td>Handle Questions</td>
</tr>
</tbody>
</table>
About Computational Genomics (BIOL350) and Genomics (BIOL650)

Grading Galaxy Work

For a given question:

- Is the answer reported correct?
- Does the corresponding Galaxy history exist?
- Does the corresponding Galaxy history contain the corresponding individual processes?
- Are those processes correct?
About Computational Genomics (BIOL350) and Genomics (BIOL650)

Grading

Grading Galaxy Work

<table>
<thead>
<tr>
<th>Number of Students</th>
<th>Number of Questions</th>
<th>Number of Processes/Question</th>
<th>Number of Processes to Grade</th>
<th>Times Six Exams/Semester</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>30</td>
</tr>
<tr>
<td>50</td>
<td>1</td>
<td>5</td>
<td>250</td>
<td>1500</td>
</tr>
<tr>
<td>50</td>
<td>10</td>
<td>5</td>
<td>2500</td>
<td>15000</td>
</tr>
<tr>
<td>65</td>
<td>10</td>
<td>5</td>
<td>3250</td>
<td>19500</td>
</tr>
</tbody>
</table>

- Grading is Hard because there are different ways to solve the same problem
- Scripting grading is possible but requires files to have consistent names! (…which requires students to name files correctly, which assumes that all students will follow instructions…)
About Digital Biology (BIOL647)

Command-line Driven

Internet

Students

VPN

GitHub

Supercomputer

Virtual Machines

CyVerse
About Digital Biology (BIOL647)

Module 01 | Module 02 | Module 03
Virtual Machines (Jetstream) | Supercomputer (Ada / Grace) |
Master the Command-line | Introduction to Supercomputers |
Basic GNU/Linux/Unix Introduction | Basic Scripting | Advanced Scripting

Jetstream work is supported by XSEDE
Grant Nº BIO210035
Funded by NSF ACI-1445604, Awarded to XSEDE
About Digital Biology (BIOL647)

Module 01

GIT/GITHUB and Version Control

GIT/GITHUB

Emacs/Org-Mode

SSH and Data Transfer

User01
User02
User03

User02 Secret
User02 Public

User01 Secret
User01 Public

User03 Secret
User03 Public

Wolf00
Wolf01
About Digital Biology (BIOL647)

Module 02

Unix101
Comprehensive Introduction to GNU/Linux/Unix

Unix102
Advanced Text manipulations, Software Compilation, and Conda/Bioconda installation and use

Unix103
Introduction to Regular Expressions, GREP, SED, and AWK

Unix104
Fundamentals of Scripting
About Digital Biology (BIOL647)

Module 03

Experimental Design
- Control group
- Treatment group

Introduction to and Scripting of Mapping and Alignment

#!/usr/bin/env bash
while getopts :i:n:f:o:r:t:v option; do
  case $option in
  i) index=$OPTARG;;
  n) name=$OPTARG;;
  r) dir=$OPTARG;;
  a) cores=$OPTARG;;
  v) echo "I do not know what $OPTARG is!";;
  esac
esac

funBwa(){
  bwa \
  -t $1 \
  -R $2 \
  -S $3 \
  -Q $4 \
  > $5;
}

NCBI Databases and NGS Quality Control

Introduction to and Scripting of Transcriptome Mapping, Assembling and Quantification (Supercomputer)

Steps 1 and 2
- Reference annotation
- RNA-seq reads

Steps 3
- HISAT
- StringTie
- Read alignments
- Assembled transcripts

Step 4
- StringTie -merge
- Merged transcripts

Step 5
- gffcompare
- Transcript statistics

Step 6
- No transcript assembly
- String Tie -eB
- Read coverage tables

Steps 7-21
- Ballgown
- Plots and differential expression tables
Who Takes BIOL350, BIOL650, and BIOL647?

- BIOL350:
  - Biology: 56%
  - Other: 45%

- BIOL650:
  - Biology: 10%
  - Other: 90%

- BIOL647:
  - Biology: 12%
  - Other: 88%
About Computational Genomics (BIOL350) and Genomics (BIOL650)

Problems Encountered

- Galaxy is intrinsically collaborative, which makes it hard to “isolate” students contributions and/or interactions.
- Because we are using the same Galaxy instantiation for Exams and Teaching, we cannot stop students from continuing processing Exam-related data, once a given exam is completed.
- Currently, we cannot unmistakably document and isolate a given student’s work on a given exam.
- Other Supercomputer users sometimes submit unreal amount of jobs to Ada/Terra while an exam is in progress.

Potential Solutions

- To have the ability to float one Galaxy instantiation/student/exam.
- Dedicated nodes for Galaxy during exams.

Other Observations/Questions

- To access the Supercomputer students need to apply for a Supercomputer account.
  After the semester is over, the majority of students will not be using their Supercomputer account again.
  Is this an unnecessary overhead on the Supercomputer facility?

- Should we have a temporary Supercomputer accounts for teaching?

- We urgently need to implement the local deployment of virtual machines and, ideally, to have those machines outside the firewall.
Acknowledgements

This talk is dedicated to the memory of

Dr. Jim Hu

My friend and colleague

With whom I developed and taught Genomics for many, many years

Jim’s constant sense of humor, and his fearless attitude towards implementing new ideas was the driving force behind these courses
I cannot thank enough the Texas A&M High Performance Research Computing facility for their unwavering, constant, support and continuous help.

Texas A&M High Performance Research Computing is one of the jewels of this institution.

Special Thanks go to:

Michael Dickens
Lisa Perez
Francis Dang
Mark Huang
Keith E. Jackson
Robin Burns
Honggao Liu