Outline

• Using XSEDE to solve large scale Bioinformatics problems. Tools for teaching Bioinformatics. What can be done on a supercomputer that cannot be done on a workstation.
• What Is XSEDE
• Overview of XSEDE Bioinformatics Resources:
  Supercomputers. Visualization. Software. Gateways. How to access these resources
• De Novo Transcriptome Assembly: Trinity
• Example of running Trinity through Galaxy
• Structural Bioinformatics
• Protein DataBank
• Visualization: VMD
• Run a Molecular Dynamics Simulation
• This lecture is on-line at http://hpcuniversity.org/trainingMaterials/202
After attending this tutorial...
You will know what Resources are Available Through XSEDE for Bioinformatics for either research or educational purposes.

During this tutorial you will:
• Run a Trinity program to assemble a transcriptome
• Run a Molecular Dynamics simulation
What is XSEDE?

• Consists of supercomputers, high-end visualization, data analysis and storage around the country. Funded by NSF. NCSA, NICS, PSC, TACC, SDSC, Univ. of Virginia, Shodor Education Foundation, SURA, Univ. of Chicago, Indiana U., Purdue, Cornell, Ohio State, UC-Berkeley, Rice, and NCAR

• Scientists and Engineers around the world use XSEDE resources and services: supercomputers, collections of data, help services.

• Accessible, at no cost to them, to a variety of scientists, including those new to supercomputing, and without specialized programming skills.
Who can benefit from access to XSEDE?

- Scientists working in:
  - data-intensive fields
  - Requiring arithmetic speed
  - Large memory
  - Large amount of data
  - Parameter sweep
  - Teaching a class

- Enables researchers who have outgrown their own computers and campus clusters to graduate to supercomputing with a minimum of additional effort.
How Can XSEDE Help?

- Massively parallel clusters (total 17,652 nodes, 226,016 processor cores)
- Large shared-memory nodes (total 451 TB memory)
- Parallel file systems (total 5,641 TB disk storage)
- Fast networking (10-GigE connection)
- Very large databases (several databases at the TB scale)
- Efficient data movement tools (globus online, uberftp)
- Consulting services provided by experts (expertise on OpenMP, workflow, science gateway, scientific database, and visualization)
## Compute Resources

<table>
<thead>
<tr>
<th>Name</th>
<th>Status</th>
<th>CPUs</th>
<th>Peak TFlops</th>
<th>Utilization</th>
<th>Running Jobs</th>
<th>Queued Jobs</th>
<th>Other Jobs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stampede</td>
<td>Healthy</td>
<td>102400</td>
<td>9600.0</td>
<td>93%</td>
<td>854</td>
<td>37</td>
<td>0</td>
</tr>
<tr>
<td>SuperMIC</td>
<td>Healthy</td>
<td>7200</td>
<td>925.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gordon Compute</td>
<td>Healthy</td>
<td>16384</td>
<td>341.0</td>
<td>91%</td>
<td>395</td>
<td>7118</td>
<td>109</td>
</tr>
<tr>
<td>Darter</td>
<td>Healthy</td>
<td>23168</td>
<td>248.9</td>
<td>92%</td>
<td>28</td>
<td>17</td>
<td>7</td>
</tr>
</tbody>
</table>
## Visualization Resources

<table>
<thead>
<tr>
<th>Name</th>
<th>CPUs</th>
<th>Peak TFlops</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maverick</td>
<td>2640</td>
<td>59.1</td>
</tr>
<tr>
<td>Nautilus</td>
<td>1024</td>
<td>8.2</td>
</tr>
</tbody>
</table>

## Storage Resources

<table>
<thead>
<tr>
<th>Name</th>
<th>File Space (TB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TACC Long-term tape Archival Storage (Ranch)</td>
<td>61440.0</td>
</tr>
<tr>
<td>NICS Long-term tape Archival Storage (HPSS)</td>
<td>17000.0</td>
</tr>
<tr>
<td>TACC Long-term Storage (Wrangler Storage)</td>
<td>10000.0</td>
</tr>
<tr>
<td>PSC Persistent disk storage (Data SuperCell)</td>
<td>4000.0</td>
</tr>
<tr>
<td>SDSC Medium-term disk storage (Data Oasis)</td>
<td>4000.0</td>
</tr>
</tbody>
</table>
Welcome to Bridges, a new concept in HPC - a system designed to support familiar, convenient interfaces and interactions for both traditional and non-traditional HPC users. It is a richly connected set of interacting systems offering a flexible mix of gateways, Hadoop and Spark ecosystems, batch processing and interactivity, based on current demand. Bridges will include:

- Compute nodes with coherent shared memory ranging from 128GB to 12TB per node
- Large-memory GPU nodes to accelerate algorithms such as deep learning
- Database nodes supporting applications varying from gateways to real-time data assimilation to large document and XML databases
- Webserver nodes to host gateways and community datasets
- Data transfer nodes with 10 GigE connections to enable data movement between Bridges and XSEDE, campuses, instruments and other advanced cyberinfrastructure

We hope the following information will help you understand what Bridges is and how you can use this unique resource in your research. If you would like to hear more about Bridges' capabilities or discuss how you can take advantage of Bridges, call us at 412-268-4960 or
Bioinformatics

• Bioinformatics is the application of computational techniques to analyze the information associated with biomolecules.

• Many of the biology projects now generate a large amount of data.

• These large amounts of data means that many of the challenges in biology have become challenges in computing.
Available Software

- Most software can be installed upon request. • We can install software that is not currently available.

Determine what computational/data resources are best for you and software needs can usually be addressed (assuming it is open, or you have a license).

- Send questions/requests to help@xsede.org
Bioinformatics Software Already Installed on XSEDE Supercomputers

- **Genome Assembly and Transcriptome Assembly:**
  - ALLPATHS-LG: *whole genome shot-gun assembler*
  - Oases: *de novo transcriptome assembler*
  - Bamtools: *toolkit for reading, writing, and manipulating BAM files*
  - Pysam: *reads SAM files*
  - Trans-ABySS: *analyzes ABySS-assembled contigs from shotgun transcriptome data.*
- **Soapdenovo, Trinity, Velvet, AbySS, Newbler**
- TopHat Cufflinks
- **Alignment Algorithms**
  - BWA, BOWTIE, SHRIMP, TopHAT
- **Molecular Dynamics**
  - AMBER, CHARMM, GROMACS, NAMD
- **Quantum Chemistry:** Gaussian
XSEDE Resources

- Extended Collaborative Support
- NIP
- Campus Champions
XSEDE RESOURCES: Services:
1. Extended Collaborative Support (ECS)

- Can solicit ECS support for help:
- Porting applications to new resources
- Providing help for portal and gateway development
- Implementing algorithmic enhancements
- Implementing parallel math libraries
- Improving scalability of codes to higher processor counts
- Optimizing codes to efficiently utilize specific resources
- Assisting with visualization, workflow, data analysis, and data transfer
- Inquire at help@xsede.org
Campus Champions:
Campus representatives. Easiest way to get started is to contact your local campus champion:
https://www.xsede.org/campus-champions
https://www.xsede.org/web/guest/current-champions

Southern University
and A&M College,
Baton Rouge,
Shizhong Yang and
Rachel Vincent-Finley
Science Gateways enables you to access XSEDE resources through web interfaces, providing ease-of-use and community specific tools & resources.
We will practice using Gateways later on so that you can get hands-on experience.
Gateways
Phylogenetic Research.
The CIPRES gateway lets biologists run parallel versions of BEAST, GARLI, MrBayes, RAxML, & MAFFT via a browser interface on SDSC’s Trestles supercomputer, an XSEDE resource
• Hosts optimized versions of RAxML and MrBayes (superior parallel scaling).
How to access XSEDE resources

An XSEDE allocation provides access to computing, visualization, and/or storage resources as well as extended support services at XSEDE service provider (SP) sites. An allocation is allotted to a researcher who serves as the principal investigator (PI) of an approved project.

An account is the specific method through which an individual (or community, in the case of science gateways) logs in to a resource to utilize the allocation.

Three types of allocations:
- Startup
- Education
- Research
Startup allocation

- Requests are reviewed continually throughout the year. Fastest way of getting started on XSEDE. Recommended for all new XSEDE users. Startup allocations are for one year, after which, the PI should request a Research allocation. A PI may have only one active Startup allocation at a time.
- Available to faculty members and postdoctoral researchers at U.S. academic research institutions, as well as to K-12 teachers for classroom use.
- Appropriate uses for Startup allocations include:
  - Application development by principal investigators
  - Experimentation on XSEDE platforms
  - Use of XSEDE systems for classroom instruction
  - Developing a science gateway
Education Allocation

• Training classes at a University that has specific begin and end dates. Support classroom instruction. Registration number, class description and known timeframe.
Startup and Education Allocations

- Are easy to request with fast approval time:

- A CV and an Abstract. Can be requested at any time.

- Startup allocations are appropriate for code development or application testing.

- Education allocations: describe the work that the student will be performing. (Approx. number of students and approx. computational resource requirement per student).

- Examples of well written proposals are on-line at the portal:
- https://portal.xsede.org/allocations-overview#writing-startupeducation
Research allocation

• May be requested for any compute, visualization, or storage resource.
• Four deadlines per year.
• Need to provide more information to justify efficient use of resources.
To submit an allocation request:
Click on Allocations, Submit/Review Request
Easiest way:

• Contact a campus champion:

https://www.xsede.org/web/guest/current-champions

No campus champion at SUNO yet,

Southern University and A&M College, Baton Rouge:

Shizhong Yang and Rachel Vincent-Finley

• Or send email to help@xsede.org
Introduction to Transcriptome Assembly

• The human genome is made up of DNA, the molecule that contains the instructions needed to build and maintain cells. These instructions are spelled out in the form of "base pairs" of four different chemicals, organized into 20,000 to 25,000 genes. For the instructions to be carried out, DNA must be "read" and transcribed - in other words, copied - into RNA. These gene readouts are called transcripts, and a transcriptome is a collection of all the gene readouts present in a cell.

• What can a transcriptome tell us?
• An RNA sequence mirrors the sequence of the DNA from which it was transcribed. Consequently, by analyzing the entire collection of RNA sequences in a cell (the transcriptome) researchers can determine when and where each gene is turned on or off in the cells and tissues of an organism.
• Depending on the technique used, it is often possible to count the number of transcripts to determine the amount of gene activity - also called gene expression - in a certain cell or tissue type.
• In humans and other organisms, nearly every cell contains the same genes, but different cells show different patterns of gene expression. These differences are responsible for the many different properties and behaviors of various cells and tissues, both in health and disease.
Introduction to Transcriptome (continued)

• By collecting and comparing transcriptomes of different types of cells, researchers can gain a deeper understanding of what constitutes a specific cell type, how that type of cell normally functions, and how changes in the normal level of gene activity may reflect or contribute to disease. In addition, transcriptomes may enable researchers to generate a comprehensive, genome-wide picture of what genes are active in which cells.

• **How can transcriptome data be used to explore gene function?**
  • The function of most genes is not yet known. A search of a transcriptome database can give researchers a list of all the tissues in which a gene is expressed, providing clues about its possible function.
  • For example, if the transcriptome database shows that an unknown gene's expression levels are dramatically higher in cancer cells than in healthy cells, the unknown gene may play a role in cell growth. Or if an unknown gene is expressed in fat tissue but not in bone or muscle tissue, the unknown gene may be involved in fat storage or metabolism. In both instances, the transcriptome data give researchers a good place to start searching for a newly found gene's function.
The National Human Genome Research Institute (NHGRI), which is part of the National Institutes of Health (NIH), has participated in two projects that created transcriptome resources for researchers around the world—the Mammalian Gene Collection initiative and the Mouse Transcriptome Project.

The Mammalian Gene Collection initiative built a free, public library of human, mouse, and rat mRNA sequences. The project was led by NHGRI and the National Cancer Institute (NCI), also part of NIH. The mouse and rat are important models with which to study human biology.

The Mouse Transcriptome Project was an NIH-supported initiative that generated a free, public database of gene transcripts for many mouse tissues. These tissue-specific gene expression data, which are mapped to the mouse genome, are available in a searchable format in the Mouse Reference Transcriptome Database.

Several other transcriptome resources exist, including those in NIH programs such as the Genotype-Tissue Expression Project (GTEx) and the Encyclopedia of DNA Elements (ENCODE). GTEx is creating a catalog of human gene expression in a variety of different tissues. ENCODE researchers aim to characterize and understand the working parts of the genome, including the transcriptome. Both Novartis and the European Molecular Biology Laboratory have well-established gene expression databases.
RNAseq

• Provide picture of RNA presence being expressed in an organism at a given moment in time
  – Gives us the ability to reconstruct the mRNA transcripts.
  – After transcript reconstruction, we can determine:
    • What products are being produced
    • Quantify patterns of expression
    • Capture alternative splicing
  – Can capture MicroRNAs (miRNAs) too, though analysis methods are different than discussed in this talk.
Methods of Assembling Reads

• Reference Assembly/Mapping
  – Using an already-assembled annotated genome as a guide to map reads against.
  – Tophat + Cufflinks

• De Novo Assembly. Memory and CPU time intensive.
  – Joining together reads into a larger sequence without relying on an already-assembled sequence of a related organism.
  – Trinity
  – Oases
## Requirements

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Mapping</th>
<th>Denovo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality reference genome</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Splice sites known</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Minimum coverage (High abundance transcript recovery)</td>
<td>10x</td>
<td>30x</td>
</tr>
<tr>
<td>Minimum coverage (Low abundance transcript recovery)</td>
<td>30x or more</td>
<td>100x or more</td>
</tr>
<tr>
<td>RAM Memory</td>
<td>Moderate</td>
<td>Substantial</td>
</tr>
<tr>
<td>CPU Time</td>
<td>Moderate</td>
<td>Substantial</td>
</tr>
</tbody>
</table>
De Novo Transcriptome Assembly

- Recent advances in sequencing technologies and assembly algorithms have facilitated the reconstruction of entire transcriptomes even without a reference genome.

- Comprehensive study of the transcriptome: primary sequence
- Abundance
What if I could:
• Assemble a 20 Gbp genome?
• Assemble a terabase of metagenomic data?
• Assemble 30 transcriptomes for 10 different primate species?
• Determine variants across 10,000 whole genomes from autism-affected families?

Transcriptome Assembly

• Involves piecing together short, low-quality reads.
• Very large dataset (Gigabases to Terabases)
• Requires large memories and parallel algorithms.
• Transcriptome assembly from billions of RNA-seq reads poses an informatic challenge.
Assembly analogy (simplified)

D. Leland Taylor, 2012, “PHAST”, Davidson College
Trinity (Installed on Bridges at PSC)

• Assembler for Short Read Transcriptome Data
  – From the Broad Institute

• Resource Intensity:
  – 1GB RAM per 1M pairs of Illumina reads.
  – 1/2 - 1 hour per million pairs of reads

  – Input to Trinity: Read out fragments from the sequencer in FASTQ format.
Trinity Components

Inchworm

Chrysalis

Butterfly
Inchworm

Determine all 25-mers

Remove error containing 25-mers

Begin contig by selecting most frequent 25-mer, excluding low complexity and singleton 25-mers

Greedy contig extension in each direction

Select next most abundant 25-mer
Inchworm: Greedy Contig Extension Strategy

<table>
<thead>
<tr>
<th>Count</th>
<th>Read</th>
<th>Count</th>
<th>Read</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Next Step: Repeat procedure in opposite direction from TCA

Greedy Path: TCACA
Inchworm

Assembles the read data set by
By greedily searching for paths.

Result is a collection of linear contigs,
with each k-mer represented only once

Grabherr et al., Nat. Biotechnol.
2011, 29(7), 644-652.
Chrysalis

Group contigs into connected components
- 24 bases must overlap and
- N reads must equally span each junction side

Build de Bruijn Graph

Assign reads to components,
using largest number of 25-mer matches
Chrysalis pools contigs
If they share at least one k-1 mer and builds individual de Bruijn graphs.

Butterfly

Graph Simplification:
1) Merge nodes into linear paths
2) Prune edges with minimal support (likely sequencing errors)

Plausible Path Discovery:
Identify paths in graph supported by reads and read pairs
Butterfly

Compacts linear paths from each de Bruijn graph and output linear sequences.

• Remove adapters
  • Quality filter
  • Remove Poly A/T
  • Remove mtDNA, rRNA
  • Convert to fasta (performed in-house)

~1.4 Billion RNA-seq reads (~200GB)

Inchworm
Typical Run time: 100 hours
Cores needed: 64

Chrysalis (Run on RAM disk)
Typical Run time: 400 hours
Cores needed: 128 cores

Quantify Graph & Butterfly (Run on RAM disk)
Typical run time: 50 hours
Cores needed: 64 cores
Reads (per sample) → Abundance estimation → Identify differentially expressed transcripts → Expression patterns, transcript clusters 

Combine reads → Assembled transcripts (all samples) → De novo assembly → Assembled transcripts → Identify coding regions 

Normalization?
Two ways of running Trinity:

## Galaxy vs. Writing Your Own PBS Scripts

<table>
<thead>
<tr>
<th>Item Evaluated</th>
<th>Galaxy</th>
<th>PBS Scripts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ease of Use</td>
<td>Intuitive for beginners</td>
<td>More difficult to learn</td>
</tr>
<tr>
<td>Amount of tools</td>
<td>Limited</td>
<td>Any tool that runs on Unix</td>
</tr>
<tr>
<td>Variety of tool options</td>
<td>Limited</td>
<td>Any option that the tool accepts</td>
</tr>
<tr>
<td>Length, size and configuration of run</td>
<td>Limited</td>
<td>Any valid configuration, length, size</td>
</tr>
<tr>
<td>Workflow log</td>
<td>Entire analysis</td>
<td>Job level log only</td>
</tr>
<tr>
<td>Style of use</td>
<td>Interactive, web page</td>
<td>Batch, command line</td>
</tr>
</tbody>
</table>
Galaxy

- Galaxy is an open-source web tool (portal) that organizes bioinformatics work-flows.
- Provides methods to:
  - Record workflows (history)
  - Share common sequence data (shared data libraries)
  - Import and use (private) sequence data
  - Share results with others (shared histories)
  - Integrate new tools into the interface
  - Run long analyses using batch queues
With a web browser, go to Galaxy page on XSEDE: dxcgalaxy.psc.edu
Go to:

- dxcgalaxy.psc.edu
- Login using your Galaxy
- username and password
Successful login view:

Left Panel: Tools

Central Panel: Viewing area for Files, History and Tools

Right Panel: History
Galaxy History Names

- A Galaxy history is a recording (log) of what you did
- Use a descriptive name to clarify your history
- To change name: 1) click on old name; 2) delete old name; 3) enter new name; 4) hit ENTER key.
Galaxy History Options

- To view options, click on the triangle next to the Options menu on the history panel.
On the right hand panel: Click on Unnamed History and write a name of your choice for the history.
Select all the files by clicking on the box at right, Import to current history → Go.
Click on Galaxy symbol to go back
## Data Library “Workshop”

<table>
<thead>
<tr>
<th>Name</th>
<th>Message</th>
<th>Data type</th>
<th>Date uploaded</th>
<th>File size</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL.10k.left.fq</td>
<td>None</td>
<td>fastq</td>
<td>Mon Oct 12 18:07:08 2015 (UTC)</td>
<td>6.7 MB</td>
</tr>
<tr>
<td>ALL.10k.right.fq</td>
<td>None</td>
<td>fastq</td>
<td>Mon Oct 12 18:07:08 2015 (UTC)</td>
<td>6.7 MB</td>
</tr>
<tr>
<td>Sp_ds.10k.left.fq</td>
<td>None</td>
<td>fastq</td>
<td>Mon Oct 12 18:07:08 2015 (UTC)</td>
<td>1.7 MB</td>
</tr>
<tr>
<td>Sp_ds.10k.right.fq</td>
<td>None</td>
<td>fastq</td>
<td>Mon Oct 12 18:07:08 2015 (UTC)</td>
<td>1.7 MB</td>
</tr>
<tr>
<td>Sp_hs.10k.left.fq</td>
<td>None</td>
<td>fastq</td>
<td>Mon Oct 12 18:07:08 2015 (UTC)</td>
<td>1.7 MB</td>
</tr>
<tr>
<td>Sp_hs.10k.right.fq</td>
<td>None</td>
<td>fastq</td>
<td>Mon Oct 12 18:07:09 2015 (UTC)</td>
<td>1.7 MB</td>
</tr>
<tr>
<td>Sp_log.10k.left.fq</td>
<td>None</td>
<td>fastq</td>
<td>Mon Oct 12 18:07:09 2015 (UTC)</td>
<td>1.7 MB</td>
</tr>
<tr>
<td>Sp_log.10k.right.fq</td>
<td>None</td>
<td>fastq</td>
<td>Mon Oct 12 18:07:09 2015 (UTC)</td>
<td>1.7 MB</td>
</tr>
<tr>
<td>Sp_plat.10k.left.fq</td>
<td>None</td>
<td>fastq</td>
<td>Mon Oct 12 18:07:10 2015 (UTC)</td>
<td>1.7 MB</td>
</tr>
<tr>
<td>Sp_plat.10k.right.fq</td>
<td>None</td>
<td>fastq</td>
<td>Mon Oct 12 18:07:10 2015 (UTC)</td>
<td>1.7 MB</td>
</tr>
</tbody>
</table>

Für selected datasets: Import to current history  Go

**TIP:** You can download individual library datasets by selecting “Download this dataset” from the context menu (triangle) next to each dataset’s name.

**TIP:** Several compression options are available for downloading multiple library datasets simultaneously:

- gzip: Recommended for fast network connections
- bzip2: Recommended for slower network connections (smaller size but takes longer to compress)
- zip: Not recommended but is provided as an option for those who cannot open the above formats.
Data that we just input into Trinity

RNA-Seq data from *Schizosaccharomyces pombe* grown in four conditions: log growth (log), plateau phase (plat), diauxic shift (ds), and heat shock (hs), each with 1M Illumina paired-end strand-specific RNA-Seq data total of 4M paired-end reads.
Trinity: NGS RNA Seq → Trinity
Step 1: Run Trinity To Assemble Transcripts:

- From the Galaxy Tools panel (on the left) click on Trinity: NGS RNA Seq, and then on Trinity.
- The center panel will now contain a screen soliciting parameters for the tool. Under the JM: box, select 10G.
- Under the CPU: box, select 2
- In the Paired or Single-end data?: box, select Paired
- In the Left/Forward strand reads: box, select the left hand read set containing all four conditions: ALL.10k.left.fq
- In the Right/Forward strand reads: box, select the right hand sequence read set containing all four conditions ALL.10k.right.fq
- In the Strand Specific Library: box, select RF.
- In the Group Pairs Distance: box, select 500.
- In the Path Reinforcement Distance: box, select 75.
- In the Use Additional Params?: box, select No.
- (DO NOT CLICK ON THE EXECUTE BUTTON)
Pretend you have run this first step.

Shared Data → Published Histories → OctoberWorkshop2

Has the three output files, (FILES 11, 12 and 13) Assembled transcriptome, timings and log.
Click on Switch to this History (Make this History your Current History)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: ALL 10k.left.fq</td>
<td></td>
</tr>
<tr>
<td>2: ALL 10k.right.fq</td>
<td></td>
</tr>
<tr>
<td>3: Sp_ds.10k.left.fq</td>
<td></td>
</tr>
<tr>
<td>4: Sp_ds.10k.right.fq</td>
<td></td>
</tr>
<tr>
<td>5: Sp_hs.10k.left.fq</td>
<td></td>
</tr>
<tr>
<td>6: Sp_hs.10k.right.fq</td>
<td></td>
</tr>
<tr>
<td>7: Sp_log.10k.left.fq</td>
<td></td>
</tr>
<tr>
<td>8: Sp_log.10k.right.fq</td>
<td></td>
</tr>
<tr>
<td>9: Sp_plat.10k.left.fq</td>
<td></td>
</tr>
<tr>
<td>10: Sp_plat.10k.right.fq</td>
<td></td>
</tr>
<tr>
<td>11: Trinity on data 1 and data 2: log</td>
<td></td>
</tr>
<tr>
<td>12: Trinity on data 1 and data 2: Assembled Transcripts</td>
<td></td>
</tr>
<tr>
<td>13: Trinity on data 1 and data 2: Trinity Timings</td>
<td></td>
</tr>
<tr>
<td>14: align_and_estimate_abundance on data 4, data 12, and data 3: lo</td>
<td></td>
</tr>
</tbody>
</table>
1. Trinity Assemble Transcripts

- When Trinity completes, it will create a *Trinity.fasta* output file.
- Trinity groups transcripts into clusters based on shared sequence content. Such a transcript cluster is very loosely referred to as a *gene*. This information is encoded in the Trinity fasta accession. An example Fasta entry for one of the transcripts is formatted like so:

```
>c115_g5_i1 len=247 path=[31015:0-148 23018:149-246]
AATCTTTTTTGGTATTGGCAGTACTGTGCTCTGGGTAGTGATTAGGGCAAAAGAAGACAC
ACAATAAAGAACCAGGTGTAGACGTCAGCAAGTCAAGGCCTTGGTTCTCAGCAGACAGA
AGACAGCCCTTCTCAATCCTCATCCCTTCCTCCCTGAACAGACATGTCTTCTGCAAGCTTCTC
CAAGTCAGTTGTTTCAACAGGAACATCATCAGAATAATTTGAAATTATGATTAGTATCTGA
TAAAGCA
```

- The accession encodes the Trinity *gene* and *isoform* information. In the example above, the accession *c115_g5_i1* indicates Trinity read cluster *c115*, gene *g5*, and isoform *i1*. Because a given run of trinity involves many many clusters of reads, each of which are assembled separately, and because the *gene* numberings are unique within a given processed read cluster, the *gene* identifier should be considered an aggregate of the read cluster and corresponding gene identifier, which in this case would be *c115_g5*.
- So, in summary, the above example corresponds to *gene id: c115_g5* encoding *isoform id: c115_g5_i1*. 
Step 2: Align and Estimate Abundance

From the Galaxy Tools panel (on the left) click on align_and_estimate_abundance listed under Expression Analysis. The center panel will now contain a screen soliciting parameters for the tool:

- Under the transcripts box, select the Trinity Assembled Transcripts produced in the last step.
- Transcript target type: Trinity
- Trinity on Data 1 and 2 Assembled Transcripts (12)
- Abundance Estimation Method: RSEM
- Alignment Method: bowtie
- In the Paired or Single-end data?: box, select Paired.
- In the Left/Forward strand reads: box, select one of the data sets: Sp_plat.10k.left.fq
- In the Right/Forward strand reads: box, select the right hand sequence for the same data set: Sp_plat.10k.right.fq
- In the Strand Specific Library: box, select RF.
Align and Estimate Abundance:

```
align_and_estimate_abundance Align and estimate transcript abundances (trinity/2014-04-13p1 and later) (Galaxy Tool Version 0.0.1)

--transcripts
    12: Trinity on data 1 and data 2: Assembled Transcripts

transcript target type
    Trinity
    see bottom

Abundance Estimation Method
    RSEM
    see bottom

Alignment Method
    bowtie
    see bottom

Paired or Single-end data?
    Paired

--left
    3: Sp_ds.10k.left.fq

Left/Forward strand reads

--right
    4: Sp_ds.10k.right.fq

Right/Reverse strand reads

--SS_lib_type
    RF

Strand-specific Library Type
```
This step needs to be repeated three more times, for each experimental condition. **Output: three files:** Isoform counts, Gene Counts, BAM file and log file. (Isoforms refers to different versions of protein from the same gene)
<table>
<thead>
<tr>
<th>transcript_id</th>
<th>gene_id</th>
<th>length</th>
<th>effective_length</th>
<th>expected_count</th>
<th>TPM</th>
<th>FPKM</th>
<th>IsoPct</th>
</tr>
</thead>
<tbody>
<tr>
<td>TR100</td>
<td>c0_g1_i1</td>
<td>TR100</td>
<td>c0_g1_i1</td>
<td>296</td>
<td>36.17</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>TR101</td>
<td>c0_g1_i1</td>
<td>TR101</td>
<td>c0_g1_i1</td>
<td>575</td>
<td>299.55</td>
<td>1.00</td>
<td>200.43</td>
</tr>
<tr>
<td>TR101</td>
<td>c1_g1_i1</td>
<td>TR101</td>
<td>c1_g1_i1</td>
<td>264</td>
<td>17.58</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>TR101</td>
<td>c2_g1_i1</td>
<td>TR101</td>
<td>c2_g1_i1</td>
<td>337</td>
<td>67.35</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>TR101</td>
<td>c3_g1_i1</td>
<td>TR101</td>
<td>c3_g1_i1</td>
<td>400</td>
<td>124.82</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>TR102</td>
<td>c0_g1_i1</td>
<td>TR102</td>
<td>c0_g1_i1</td>
<td>458</td>
<td>182.55</td>
<td>4.00</td>
<td>1315.56</td>
</tr>
<tr>
<td>TR103</td>
<td>c0_g1_i1</td>
<td>TR103</td>
<td>c0_g1_i1</td>
<td>526</td>
<td>250.55</td>
<td>14.00</td>
<td>3354.79</td>
</tr>
<tr>
<td>TR103</td>
<td>c0_g1_i2</td>
<td>TR103</td>
<td>c0_g1_i2</td>
<td>526</td>
<td>250.55</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>TR104</td>
<td>c0_g1_i1</td>
<td>TR104</td>
<td>c0_g1_i1</td>
<td>393</td>
<td>118.01</td>
<td>7.00</td>
<td>3561.35</td>
</tr>
<tr>
<td>TR105</td>
<td>c0_g1_i1</td>
<td>TR105</td>
<td>c0_g1_i1</td>
<td>1222</td>
<td>946.55</td>
<td>52.00</td>
<td>3298.32</td>
</tr>
<tr>
<td>TR106</td>
<td>c0_g1_i1</td>
<td>TR106</td>
<td>c0_g1_i1</td>
<td>747</td>
<td>471.55</td>
<td>7.00</td>
<td>891.22</td>
</tr>
<tr>
<td>TR106</td>
<td>c0_g2_i1</td>
<td>TR106</td>
<td>c0_g2_i1</td>
<td>774</td>
<td>498.55</td>
<td>0.00</td>
<td>0.04</td>
</tr>
<tr>
<td>TR106</td>
<td>c1_g1_i1</td>
<td>TR106</td>
<td>c1_g1_i1</td>
<td>1969</td>
<td>1693.55</td>
<td>8.01</td>
<td>283.87</td>
</tr>
</tbody>
</table>
3. Abundance Estimate to Matrix

![Screenshot of Galaxy software interface showing the abundance_estimate_to_matrix tool.](image)

The image shows the Galaxy software interface with the abundance_estimate_to_matrix tool open. The tool is used to create a condition replicate file and estimate abundances from RSEM or eXpress results. The screenshot includes a table with columns for `Condition Name`, `Replicate Name`, and `Counts Matrix`, among others.
**Output**
The table cell in the i-th row and the j-th column of the table tells how many reads have been mapped to gene i in sample j.

<table>
<thead>
<tr>
<th></th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
</tr>
</thead>
<tbody>
<tr>
<td>TR193</td>
<td>c1_g1_i1</td>
<td>3.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>TR9</td>
<td>c0_g1_i1</td>
<td>14.00</td>
<td>1.00</td>
<td>15.00</td>
</tr>
<tr>
<td>TR53</td>
<td>c3_g1_i1</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>TR88</td>
<td>c0_g1_i1</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>TR138</td>
<td>c0_g1_i1</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>TR252</td>
<td>c1_g1_i1</td>
<td>19.00</td>
<td>15.00</td>
<td>0.00</td>
</tr>
<tr>
<td>TR215</td>
<td>c0_g1_i1</td>
<td>20.00</td>
<td>49.00</td>
<td>12.00</td>
</tr>
<tr>
<td>TR73</td>
<td>c0_g1_i1</td>
<td>122.00</td>
<td>1.00</td>
<td>76.00</td>
</tr>
<tr>
<td>TR316</td>
<td>c0_g1_i1</td>
<td>1.00</td>
<td>0.00</td>
<td>10.00</td>
</tr>
<tr>
<td>TR319</td>
<td>c0_g1_i1</td>
<td>8.00</td>
<td>28.00</td>
<td>10.00</td>
</tr>
</tbody>
</table>
Differential Expression Analysis

• Input Sample File: Condition Replicate file from previous step.
• Input Matrix: Condition Replicate Counts

Output: MA and Volcano plots in pdf format
Visualization for comparing transcript expression profiles between two experimental conditions.

**Volcano plot**

X-axis: FC: fold change or effect:  log (experiment/control).  Y-axis log (p value)= significance. Volcano plot gives you the significance (p. value) per reference. The lower the p value, bigger the y value and more significantly the event happens, and more likely the effect could be true. Points located at topright and topleft are probably true Differentially expressed genes. FDR= False Discovery Rate.
Heatmap: Genes (transcripts) vs samples
If everything else fails...

- Shared data ➔ Published History ➔ WorkshopCompleteAnalysis
- Switch to this History (makes this history the current one)
- By clicking on the name of the instruction
- One can run it again:
Align and estimate transcript abundances (trinity/r2014-04-13p1 and later) (Galaxy Tool Version 0.0.1)

--transcripts
12: Trinity on data 1 and data 2: Assembled Transcripts

transcript target type
Trinity

Abundance Estimation Method
RSEM

Alignment Method
bowtie

Paired or Single-end data?
Paired

--left
3: Sp_ds.10k.left.fq

Left/Forward strand reads

--right
4: Sp_ds.10k.right.fq

Right/Reverse strand reads

--SS_lib_type
RF

15: align_and_estimate_abundance
17: align_and_estimate_abundance on data
16: align_and_estimate_abundance on data
4. data 12, and data 3: BAM file
4. data 12, and data 3: Gene count results
473 lines
format: geneCounts, database: 7

Settings:
Output files:
Output file format:
Line rate: 6 (line is 64 bytes)
Lines per side: 1 (side is 64 bytes)
Offset rate: 5 (one in 32)

<table>
<thead>
<tr>
<th>DATA SOURCE</th>
<th>BIOINFORMATICS TOPICS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Macromolecular Structure And Dynamics</td>
<td>Protein Data Bank. VMD. Molecular dynamics (AMBER. NAMD)</td>
</tr>
<tr>
<td>RNAseq</td>
<td>Trinity</td>
</tr>
</tbody>
</table>
The 3D structure of proteins is mainly determined by X-ray crystallography or by nuclear magnetic resonance (NMR). In X-ray crystallography, one attempts to infer the 3D position of each of the protein’s atoms from a projection obtained by passing X-rays through a crystallized sample of that protein.

In the NMR technique, one obtains a number of matrices that express the fact that two atoms are within a certain distance. One then deduces a 3D shape from those matrices. A great advantage is that they allow one to study mobile parts of proteins, a task which cannot be done using crystals.
PDB database

- The Protein Data Bank (PDB) archives information about the 3D structures of large biological molecules, including proteins and nucleic acids.

- Understanding the shape of a molecule deduce a structure's role in human health and disease, and in drug development. The structures in the archive range from tiny proteins and bits of DNA to complex molecular machines like the ribosome.

PDB Home page:
http://www.rcsb.org/pdb/home/home.do
A Structural View of Biology

This resource is powered by the Protein Data Bank archive—information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.

The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

Take an Interactive Tour of the PDB

September Molecule of the Month

Amyloids
Structural Bioinformatics: Exercise

• 1. Download the protein Ubiquitin from the PDB

• 2. Visualize with VMD (Visual Molecular Dynamics)
1. Go to: http://www.rcsb.org/pdb/home/home.do
STRUCTURE OF UBIQUITIN REFINED AT 1.8 ANGSTROMS RESOLUTION

DOI: 10.2210/pdb1ubq/pdb

Primary Citation

Structure of ubiquitin refined at 1.8 A resolution.
Vijay-Kumar, S., Bugg, C.E., Cook, W.J.
PubMed: 3041007
Search Related Articles in PubMed

PubMed Abstract:
The crystal structure of human erythrocytic ubiquitin has been refined at 1.8 A resolution using a restrained least-squares procedure. The crystallographic R-factor for the final model is 0.176. Bond lengths and bond angles in the molecule have root-mean-square deviations from ideal values of molecules per molecule of ubiquitin are...
VMD. We have already installed it on your workstations.
• Open a terminal and type vmd
• Mouse: Rotate, Scale
• Graphics → Graphical Representations →
• Coloring Style → Secondary Structure
• Drawing → Ribbons
• Label → Bonds and then click on two atoms to get the distance
• Label → Angles and then click on three atoms to get the angle between them.
Resources for Teaching

- PDB available on the Web.
- VMD available from UIUC
- We now go to XSEDE Resources
• Crystallographic Structure: Static Representation
• Cannot explain how molecules bind or interact with each other. Or how a drug gets into the binding pocket.
A movie is worth a million pictures....

Nude descending a Staircase,
Marcel Duchamp
Molecular Dynamics

For each atom in the molecule, we need:

- Position ($r$)
- Momentum ($m + v$)
- Charge ($q$)
- Bond information (which atoms, bond angles, etc.)
To run the simulation, we need the force on each particle.

We use the gradient of the potential energy function.

Now we can find the acceleration.
What is the Potential?

A single atom will be affected by the potential energy functions of every atom in the system:

- Bonded Neighbors
- Non-Bonded Atoms (either other atoms in the same molecule, or atoms from different molecules)

\[ V(R) = E_{\text{bonded}} + E_{\text{non–bonded}} \]
Non-Bonded Atoms

There are two potential functions we need to be concerned about between non-bonded atoms:

• van der Waals Potential
• Electrostatic Potential

\[ E_{\text{non-bonded}} = E_{\text{van-der-Waals}} + E_{\text{electrostatic}} \]
The van der Waals Potential

- Atoms with no net electrostatic charge will still tend to attract each other at short distances, as long as they don’t get too close.
- Once the atoms are close enough to have overlapping electron clouds, they will repel each other with astounding force.

\[
E_{\text{Lennard–Jones}} = \sum_{\text{nonbonded pairs}} \left( \frac{A_{ik}}{r_{ik}^{12}} - \frac{C_{ik}}{r_{ik}^6} \right)
\]
The Electrostatic Potential

- Opposite Charges Attract
- Like Charges Repel
- The force of the attraction is inversely proportional to the square of the distance

\[ E_{\text{electrostatic}} = \sum_{\text{nonbonded pairs}} \frac{q_i q_k}{D r_{ik}} \]
Coulomb’s Law

\[ F = \frac{q_1 q_2}{4 \pi \epsilon_0 r^2} \]
The Non-Bonded Potential

\[ E_{\text{non-bonded}} = E_{\text{van-der-Waals}} + E_{\text{electrostatic}} \]
Bonded Atoms

There are three types of interaction between bonded atoms:

• Stretching along the bond
• Bending between bonds
• Rotating around

\[ E_{\text{bonded}} = E_{\text{bond-stretch}} + E_{\text{angle-bend}} + E_{\text{rotate-along-bond}} \]
Both the spring constant and the ideal bond length are dependent on the atoms involved.

\[ E_{\text{bond-stretch}} = \sum_{1,2 \text{ pairs}} K_b (b - b_0)^2 \]
Bond Angle Potentials

The spring constant and the ideal angle are also dependent on the chemical type of the atoms.

\[ E_{\text{bond-bend}} = \sum_{\text{angles}} K_\theta (\theta - \theta_0)^2 \]
Verlet Algorithm

• First, take a third-order Taylor step:

\[
\begin{align*}
    r(t + \Delta t) &= r(t) + v(t)\Delta t + \frac{1}{2} a(t)\Delta t^2 + \frac{1}{3!} r(t)\Delta t^3 + O(\Delta t^4) \\
    r(t - \Delta t) &= r(t) - v(t)\Delta t + \frac{1}{2} a(t)\Delta t^2 - \frac{1}{3!} r(t)\Delta t^3 + O(\Delta t^4)
\end{align*}
\]
Verlet Algorithm

• When adding the two formulas, the first and third derivatives cancel out:

\[ r(t + \Delta t) + r(t - \Delta t) = 2r(t) + a(t)\Delta t^2 + O(\Delta t^4) \]

• And we can express the next timestep in terms of the previous position and the current acceleration:

\[ r(t + \Delta t) = 2r(t) - r(t - \Delta t) + a(t)\Delta t^2 + O(\Delta t^4) \]
Periodic Boundary Conditions

- Simulate a segment of molecules in a larger solution by having repeatable regions

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIII</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VII</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Cutoff Methods

• Ideally, every atom should interact with every other atom
• This creates a force calculation algorithm of quadratic order
• We may be able to ignore atoms at large distances from each other without suffering too much loss of accuracy
Running a Molecular Dynamics Simulation using XSEDE resources

• Go back to the Portal:
• portal.xsede.org
• Click on Resources → Science Gateways
portal.xsede.org. Click on Visit Portal (on the right) for the first Gateway (Diagrid).
Choose your sign in method:

- Sign in with Purdue Career
- Sign in with your diagrid account

Create an account
Sign In. Then click on NAMD under tools.
NAMDD
By Feng (Kevin) Chen¹, Rob Campbell
1. Purdue University
Upload files, execute NAMD molecular dynamics simulation jobs and download results.

Abstract
NAMDD is a graphical interface to submit NAMD simulation jobs to High Performance Computing (HPC) Clusters. The tool is capable of reading in existing NAMD configuration files and then send it to a remote host for execution.
• Click on Lunch application and
• Accept running JAVA
How to unblock Java

- **Safari**: Click on Safari and select **Preferences**
  - Choose the **Security** option
  - Select **Allow Plug-ins**, then click on **Manage Website Settings**
  - Click on the Java item, select an option (Ask, Allow or Allow Always) from the pulldown list

  *When visiting other websites*
  - Click **Done**, then close the Safari Preferences window

- **Internet Explorer**: Click **Tools** and then **Internet Options**
  - Select the **Security** tab, and select the **Custom Level** button
  - Scroll down to **Scripting of Java applets**
  - Make sure the **Enable** radio button is checked
  - Click **OK** to save your preference

- **Firefox**: Open the Firefox browser or restart it, if it is already running
  - From the Firefox menu, select **Tools**, then click the **Add-ons** option
  - In the Add-ons Manager window, select **Plugins**
  - Click **Java (TM) Platform** plugin (Windows) or **Java Applet Plug-in** (Mac OS X) to select it
  - Check that the option selected is **Ask to Activate** or **Always Activate** or on older Firefox versions, click on the **Enable** button (if the button says **Disable**, Java is already enabled)
Do you want to trust the website "diagrid.org" to use the "Java" plug-in?

You can manage which websites you trust to use this plug-in in the Security pane of Safari preferences.

Trust  Cancel
Import input files

Click on:
Import Files
Browse
/home/diagrid/USERNAME
data
tools
namdd
XSEDE_outreach_session_New_Orleans.zip
Input files needed to run NAMD

Upload files, execute NAMD simulation jobs, download results.

Status: Ready to configure job.

Log
Files
Output
par_all27_prot_lipid.inp
ubq_wb_eq.dcd
prog.namd
ubq_ws.pdb
ubq_ws.psf

Step 1 - Import Files
Step 2 - Configure Job
Step 3 - Submit Job
Step 4 - Export Files
Input files needed to run NAMD

Upload files, execute NAMD simulation jobs, download results.

Status: Ready to configure job.

Log | Files | Output
--- | --- | ---
par_all27_prot_lipid.inp
ubq_wb_eq.dcd
prog.namd
ubq_ws.pdb
ubq_ws.psf
Input files needed to run NAMD

Upload files, execute NAMD simulation jobs, download results.

Status: Ready to configure job.

File list:
- par_all27_prot_lipid.inp
- ubq_wb_eq.dcd
- prog.namd
- ubq_ws.pdb
- ubq_ws.psf

Steps:
1. Import Files
2. Configure Job
3. Submit Job
4. Export Files
When the job completes, export files
Can look at the trajectory (dcd file) with VMD:
Load first the pdf file (as a New Molecule) and then the dcd file on top of it (not as a new Molecule).
Run Molecular Dynamics on a Supercomputer

• Start a terminal window.
• ssh train0...@login.xsede.org
• Enter Portal Password.
• gsissh comet.sdsc.xsede.org

• Copy job script and input files:
• cp /home/mmadrid/namd_input/* .

Job script: namd_comet.job

Input files: ubq_ws.pdb (Ubiquitin + water)
            ubq_ws.psf
            par_all27_prot_lipid.inp
            prog.namd
• qsub namd_comet.job

• qstat -a
Output Files

- prog16.log
- namd.job.o418176

- Can look at this files using more

To make sure there are no error messages.
Output files

- List all output files with
- `ls -ltr`
• Bring them to your computer to look at them with vmd:

• From a window on your computer:
  • sftp train0...@comet.sdsc.xsede.org
  • Cd ...
  • Mget *
Conclusions

• Have learned how to assemble a transcriptome, and how to run molecular dynamics and visualize the results.
How do I get started using XSEDE resources?

To get started using XSEDE, a researcher needs to:

- apply for an XSEDE allocation
- request to be added to an existing one.

You do either of these through the Portal.

Or

- Contact your local campus champion.
• We are here to help you...
Need help? Reporting and Tracking Issues

• portal.xsede.org → Help
  Submit ticket

• portal.xsede.org → My XSEDE → Tickets
  – Submit ticket
  – View past tickets (both open and closed)

• Can also email help@xsede.org or call 1-866-907-2383, at any hour (24/7)
Help: Submitting a ticket

Preferred method:

Helps ensure all relevant details are provided for quick routing.
Please complete the survey at:

• http://bit.ly/xsedesuno
Thanks for listening and welcome to XSEDE!
Import files, option II

(to be used only of Option 1 fails)

Go to the tool resource page

https://diagrid.org/tools/namdd

- Supporting Docs tab
- XSEDE_outreach_session_New_Orleans.zip - This will download a copy to the users local machine.
- Step 1 - Import Files
- Upload
- Browse on local machine to find file downloaded earlier.
- Upload
DiaGrid

134,245,027 jobs run to date
No Forms. No waiting. Just instant access to high-throughput computing

DiaGrid helps advance biomedical imaging for understanding and treating cancer and other diseases
Purdue's DiaGrid hub is helping advance a biomedical imaging technique paradigm

Tools
We support BLAST, R scripts, CESM, Gromacs, NAMD, BEAST and other programs used by thousands of researchers. Find a tool, click the launch

Incentives
Earn your way to VIP Status and receive more cycles and higher priority. Tell us about what you're doing and earn more cycles. Get involved by asking/answering
Click on Supporting Docs and then on XSEDE_outreach_session_New_Orleans.zip
Files will be downloaded. Then Click on Upload and Upload the files from your computer.