University of Missouri (MU) Cyberinfrastructure (CI) Plan Update
Bioinformatics and Genomics
February 11, 2016

Note: This document was prepared as part of MU’s CI Council process for updating the campus plan for research cyberinfrastructure by this group: Bob Schnabel, Scott Givan, Alan Marshall, Elizabeth (Libby) King, Tim Middelkoop and Diane Oerly. This brief document is intended to accompany the 32-page detailed analysis of the impact of high-throughput sequencing on MU research capacities. As that report documents, MU’s 2005 investment ($231,100 cost share) in next-generation sequencing has been returned many-fold. Within ten years, MU researchers have been awarded a total of 66 grants involving sequencing, totaling $87.5 million ($7.6 million directly attributable to sequence data generation and analysis) plus, another $7.9 million in grant funding proposals are pending or submitted.

Progress made since January 2013 (formation of CI Council)
The surveys of faculty and graduate students in Spring 2013, and High Performance Computing (HPC) focused survey and focus group meeting as well as MU’s first CI Day held in October, 2013 served to recognize that High Memory HPC for genomics studies was a particular need for cyberinfrastructure that was not met by the HPC equipment available on campus. Progress in general purpose research storage (GPRS) does, of course, apply to bioinformatics and genomics researchers. And, MU’s 100GB network connectivity does enable MU biology and genomics researchers to collaborate with their colleagues across the nation and around the world. While the 2014 and 2015 purchases and upgrades to ‘Lewis3’ helped address the needs for general-purpose HPC, there is only one node with sufficient memory to run the bioinformatics and genomics processes. Through collaboration and consultation with these researchers, this update to MU’s CI plan was created to address the needs of this particular subset of MU researchers.

Collaborations with faculty led to a total of $1 million in MRI and CC*IIE grants from the NSF in 2014. These have improved MU’s HPC gear and provided a CI engineer position to enabled the creation of a ‘sandbox’ that will enable Research Support Computing to work with faculty to discern more intelligent purchases including: types of memory, and the performance of various options for disk storage (OS, test file systems, I/O configurations), scaling - the optimal number of cores, the type of DIMS, open stack virtualized environment, etc.

Immediate Situation
This table provides a summary of the bioinformatics and genomics computational requirements at MU as of summer 2015. Clearly there is demand for several machines with high core count and either 0.5 or 1 TB of memory. And, some of these large memory machines need to have high-speed local disk storage. A more detailed description of the bioinformatics situation is contained as an attachment to this document.

<table>
<thead>
<tr>
<th>Analysis Type:</th>
<th># Cores</th>
<th>Ram (G)</th>
<th>Scratch (T)</th>
<th>Jobs /Batch</th>
<th>Batches</th>
<th>Length (Day)</th>
<th># Labs</th>
<th>Core Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error Correction</td>
<td>64</td>
<td>512</td>
<td>1</td>
<td>400</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>6,144,000</td>
</tr>
<tr>
<td>RNA Assembly - Big</td>
<td>64</td>
<td>1024</td>
<td>1</td>
<td>10</td>
<td>2</td>
<td>14</td>
<td>5</td>
<td>2,150,400</td>
</tr>
<tr>
<td>RNA Assembly – Small</td>
<td>64</td>
<td>512</td>
<td>1</td>
<td>100</td>
<td>2</td>
<td>0.5</td>
<td>5</td>
<td>768,000</td>
</tr>
<tr>
<td>RNA-Seq Alignment</td>
<td>20</td>
<td>128</td>
<td>1</td>
<td>40</td>
<td>50</td>
<td>0.25</td>
<td>1*</td>
<td>240,000</td>
</tr>
<tr>
<td>RNA-Seq Analysis</td>
<td>40</td>
<td>512</td>
<td>1</td>
<td>1</td>
<td>50</td>
<td>1</td>
<td>1*</td>
<td>48,000</td>
</tr>
<tr>
<td>Genome Assembly – Prokaryote</td>
<td>≥40</td>
<td>256</td>
<td>1</td>
<td>1</td>
<td>12</td>
<td>1</td>
<td>1*</td>
<td>11,520</td>
</tr>
<tr>
<td>Genome Assembly – Eukaryote</td>
<td>≥60</td>
<td>768</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>10</td>
<td>1*</td>
<td>28,800</td>
</tr>
<tr>
<td>Metagenomics</td>
<td>60</td>
<td>128</td>
<td>1</td>
<td>10</td>
<td>100</td>
<td>2</td>
<td>1*</td>
<td>2,880,000</td>
</tr>
</tbody>
</table>
Metagenomic Assembly ≥60 1024 1 1 2 10 1* 28,800
Illumina Base Calling 10 128 1 24 120 0.5 1* 345,600
DNA-Seq Alignment 40 128 1 2 10 0.5 1* 9,600
Methyl-Seq Alignment 40 128 1 12 6 1 1* 69,120
Methyl-Seq Analysis 20 256 1 4 6 1 1* 11,520
Sequence DB Search 5 128 1 50000 20 0.005 1* 600,000
Variant Detection/Analysis 40 256 1 12 12 1 1* 138,240
Genome Annotation – Prokaryote 5 256 2 50000 6 0.01 1* 360,000
Genome Annotation – Eukaryote 5 512 2 200000 1 0.01 1* 240,000
Genome Analysis ≥60 768 1 2 6 2 1* 34,560

* Jobs done by the IRCF represent a client base of over 70 labs

**Currently Available Machines:**

<table>
<thead>
<tr>
<th>Cores</th>
<th>Platform</th>
<th>RAM (G)</th>
<th># of Machines</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>Intel</td>
<td>1024</td>
<td>1</td>
<td>Lewis Cluster</td>
</tr>
<tr>
<td>24</td>
<td>AMD</td>
<td>1024</td>
<td>1*</td>
<td>IRCF BioCluster</td>
</tr>
<tr>
<td>64</td>
<td>AMD</td>
<td>512</td>
<td>4*</td>
<td>IRCF BioCluster</td>
</tr>
<tr>
<td>64</td>
<td>AMD</td>
<td>512</td>
<td>5*</td>
<td>MUgenomics Cluster</td>
</tr>
</tbody>
</table>

* Note that these are independently maintained by IRCF and Animal Sciences and are not generally available to the MU campus.

**Status:**

- Two high memory machines (512GB and 1TB) will be purchased as part of the 2014 MRI project (Shyu, PI). 3 machines, one of which will have next generation SSD technology (NVMe) for high speed local disk I/O access. These high-core, large memory, high I/O, machines will become a part of the experimental cluster, and when the three year project ends will be incorporated into general use.

- One role of MU’s Informatics Research Core Facility (IRCF) is to facilitate bioinformatics and genomics research, which includes assisting current and prospective faculty to gain access to the necessary computational hardware and software. The model for long-term sustainability involves collaborating in funded research and collecting revenue for services.

**Objectives for Bioinformatics and Genomics CI Plan Update**

**One Year Objectives**

1. Based on needs assessment conducted as part of the CI Plan update, evaluate several vendor options for a bioinformatics cluster. Identify the best performance to cost ratio, and determine a cost per unit of sequence data.

   A small number of possible vendors have been selected, and evaluation criteria have been established. Using the HPC system purchased through the National Science Foundation as well as the Lewis3 Cluster and existing data storage and networking will assure a realistic evaluation.

2. Evaluate alternatives for reliable and cost effective data storage to serve as high speed scratch space in the bioinformatics data processes.

   Scalability is critical, and advance analysis and testing will assure informed choices of how these systems can grow with demand. The combined systems of 1) computational capacity and 2) high speed scratch data storage will increase incrementally as more investor researchers come forward.
3. Prepare a management plan for the Bioinformatics Cluster. 
   Plan will address how the systems will integrate with existing HPC equipment within the 
   Informatics Research Core Facility and DoIT’s Research Computing Support Services, 
   and a business model likely similar to the ‘investor model’ first employed in the GPRS. 
   An estimated “cost per unit of sequence data” will be determined for researchers to use in 
   determining budget requirements for funding research proposals.

4. Develop a funding proposal to be presented to MU’s Office of Research for a bioinformatics 
   cluster designed to support bioinformatics and genomics researchers at the University of 
   Missouri. 
   Equipment choice will be determined based on the evaluations which were guided by the 
   needs assessment generated in the CI Plan update process. Sustainable budget will be 
   based on the business model and business plan.

Three Year Objectives
1. Continue to increase the number of collaborations with MU researchers to develop funding 
   proposals that result in equipment that not only addressed the needs of a particular project, but 
   integrate with other systems to support the broader research community.
2. Enhance the support model for researchers to assure researchers have access to the necessary 
   combination of domain area expertise and working knowledge of MU’s research computing 
   environment.
3. Review and refine the Investor Model of funding to assure it is reasonable from the investors’ 
   perspective and is resulting in financially sustainable systems.

Appendix - Impact of Sequencing Research
Lengthy (30+ pages) document describing the history and growth in sequencing and the 
impacts in grants awarded, publications, and the demand for computational capacity.