Teaching Bioinformatics data analysis using *Medicago truncatula* as a model

Vivek Krishnakumar

Session: Teaching Genetics, Genomics, Bioinformatics and Biotechnology

Plant & Animal Genome XXIV

Saturday, Jan 9th, 2016
Outline

• Background
  o Medicago genome project
  o Outreach mandate
  o Our Vision

• JCVI Plant Bioinformatics Workshop

• Community access to workshop resources

• Related Initiatives

• Summary
Medicago genome project

- *Medicago truncatula*, a close relative of alfalfa, is the preeminent model for legume genomics
- Sequencing initiated in 2003, renewed in 2006, moved to curation phase in 2009
- Funded by NSF Plant Genome awards #0321460, #0604966 and #0821966, respectively
Medicago genome project activities

- Sequencing
  - Sanger-based BAC sequencing
  - Sequence finishing/gap closure
  - NextGen sequencing (NGS) using Illumina/454
- Assembly
  - Tiling-path & genetic map based genome assembly
  - Whole Genome Shotgun (WGS) assembly
  - Optical Map based genome assembly improvement
- Annotation
  - de novo gene finding, transposon classification
  - Transcriptome based gene structural annotation
  - Transcriptome based Alternative Splicing (AS) detection
  - Gene functional annotation
- Online Databases
  - Medicago truncatula Genome Database
  - Medicago Community Annotation Portal
Outreach Mandate

NSF Award #0821966:

At the educational level, participating institutions will host visiting students in their laboratories for summer internships. In addition, annual workshops will be held to provide education in genome annotation and analysis to graduate students, postdoctoral fellows and interested faculty in the legume community.

Our Vision

- Genome and transcriptome sequencing is now commonplace, sequencing tech constantly evolving
- New methodologies and tools to analyze/visualize data continue to be developed and released
- Pressing need for researchers to keep abreast of new bioinformatics analysis techniques
- Goal:
  - Develop a comprehensive curriculum capable of covering theoretical and practical nuances of genomic data analysis, targeted towards researchers looking to hone their bioinformatics skills
JCVI Plant Bioinformatics Workshop

Background

- Annual week-long workshop
- Started in 2010 and concluded in 2014
- Open to participants within/outside the USA
- Open to university and industry participants
- Open to remotely located participants
- Fully paid for by the NSF Award (except for international travel)
- Focused on various aspects of Genomics and Bioinformatics data analysis
JCVI Plant Bioinformatics Workshop
Presentations

- Internal instructors present talks on topics deriving from their day-to-day tasks within the Plant Genomics group
  - **Linux**: Getting familiar with command line interface (CLI),
    1. learning to use command line toolkits
    2. understanding common file formats (GFF3, BED, SAM)
  - **Assembly**:
    1. genome sequencing technologies (454, Illumina, PacBio)
    2. genome assembly methods and tools (SOAP de novo, Velvet)
    3. assembly comparison tools (nucmer)
  - **Annotation**:
    1. gene finding methodologies
    2. functional annotation tools
    3. transcriptome assembly and analysis
    4. differential expression analysis
  - **Variation**:
    1. Single Nucleotide Variations (SNV) and their effects
    2. Variant analysis tools
- Guest instructors present domain specific talks: small RNA analysis (Blake Meyers, DBI), Repeat analysis (Heidrun Gundlach, MIPS), Comparative genomics (Eric Lyons, UofA/iPlant), Quantifying transcript abundance (Andrew Farmer, NCGR), Synthetic Biology (Other JCVI Researchers)
Hands-on data analysis sessions are interspersed between presentations.

Exercises are designed against real data, either generated by the Medicago project, or other published datasets.

Attendees perform all the data analysis on the command-line interface, directly on JCVI hosted computational resources.

Computational needs for remote attendees managed via cloud compute technology powered by Amazon web services.
JCVI Plant Bioinformatics Workshop

Cloud-based collaboration technologies

- Cloud-based document sharing
  - Google Drive platform
  - Presentation and hands-on material hosted as live documents
  - Content organized into logical folders
  - Content accessible after workshop completion

- Cloud-based teleconferencing
  - Cisco WebEx platform
  - Facilitates instantaneous voice and video calling
  - Share content with remote participants
  - Selective recording of talks
Setting up and testing compute, data and analysis tools within JCVI enabled estimation of resource requirements in terms of CPU, RAM and storage.

Resources replicated onto the Amazon Elastic Cloud Compute (EC2) infrastructure to build Virtual Machine (VM) image.

VM image used to spawn on-demand instances as per requirements of remote attendees.

<table>
<thead>
<tr>
<th>Resource</th>
<th>Allocation (per machine)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processing Cores</td>
<td>20 CPU</td>
</tr>
<tr>
<td>Memory (RAM)</td>
<td>40 GB</td>
</tr>
<tr>
<td>Storage</td>
<td>150 GB</td>
</tr>
</tbody>
</table>

For a total of 20 users, 4x machines allocated.
## JCVI Plant Bioinformatics Workshop Participation

<table>
<thead>
<tr>
<th></th>
<th>Undergrad &amp; Graduate Students</th>
<th>Postdocs/Scientist</th>
<th>Faculty</th>
<th>Women</th>
<th>Universities</th>
<th>Intl. Universities</th>
<th>Industries</th>
<th>Govt. Agencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workshop 2014</td>
<td>7</td>
<td>11</td>
<td>4</td>
<td>10</td>
<td>14</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Workshop 2013</td>
<td>8</td>
<td>5</td>
<td>4</td>
<td>7</td>
<td>15</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Totals</td>
<td>15</td>
<td>16</td>
<td>8</td>
<td>17</td>
<td>29</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>

Community access to workshop resources

- For posterity, complete set of workshop resources have been posted as a free-to-user Virtual Machine (VM) image available on the open-access cloud computing infrastructure, **Atmosphere**, developed and made available by **CyVerse** (formerly **iPlant Collaborative**)

- VM image: [https://atmo.iplantcollaborative.org/application/images/899](https://atmo.iplantcollaborative.org/application/images/899)

- Presentations & Hands-on exercise material: [http://j.mp/jcvi-bioinfo-workshop](http://j.mp/jcvi-bioinfo-workshop)
Community access to workshop resources

Requirements to access these resources:

- Create an iPlant account: [https://user.iplantcollaborative.org](https://user.iplantcollaborative.org)
- Request access to Atmosphere: [https://pods.iplantcollaborative.org/wiki/x/mIly](https://pods.iplantcollaborative.org/wiki/x/mIly)
- Create new instance from Workshop VM image: [https://pods.iplantcollaborative.org/wiki/x/Blm](https://pods.iplantcollaborative.org/wiki/x/Blm)
- Once instance is running, follow the SSH instructions from “Connecting to iPlant Instance” document in the Google Docs repository: [http://j.mp/jcvi-bioinfo-workshop](http://j.mp/jcvi-bioinfo-workshop)

Layout of data and tools:

```
$ ls -lht
-rw-rw-r-- 1 htang tigr 264 Jun 12 2014 rnamespace
-rw-rw-r-- 1 htang tigr 164 Jun 12 2014 assembly
-rw-rw-r-- 1 htang tigr 438 Jun 12 2014 annotation
-rw-rw-r-- 1 htang tigr 48 Jun 12 2014 doc
-rw-rw-r-- 1 htang tigr 298 Jun 12 2014 packages
-rw-rw-r-- 1 htang tigr 8 Jun 12 2014 lib
-rw-rw-r-- 1 htang tigr 122 Jun 12 2014 apps
-rw-rw-r-- 1 htang tigr 44 Jun 12 2014 snp
-rw-rw-r-- 1 htang tigr 21 Jun 12 2014 share
-rw-rw-r-- 1 htang tigr 64 Jun 12 2014 rsync.sh
-rw-rw-r-- 1 htang tigr 262 Jun 14 2014 rsync-aws.sh
-rw-rw-r-- 1 htang tigr 302 Jun 14 2014 rsync.sh
-rw-rw-r-- 1 htang tigr 17K Jul 30 2014 bin
```

Component specific layout:

```
$ ls -lht
-rw-rw-r-- 1 htang tigr 114 Jun 12 2014 trinity
-rw-rw-r-- 1 htang tigr 114 Jun 12 2014 tophat
-rw-rw-r-- 1 htang tigr 114 Jun 12 2014 rsem
-rw-rw-r-- 1 htang tigr 45 Jun 12 2014 fastqc
-rw-rw-r-- 1 htang tigr 374 Jun 12 2014 data
-rw-rw-r-- 1 htang tigr 24 Jun 12 2014 cufflinks
-rw-rw-r-- 1 htang tigr 99 Jun 12 2014 cuffcompare
-rw-rw-r-- 1 htang tigr 72 Jun 12 2014 bowtie2-build
-rw-rw-r-- 1 htang tigr 238 Jun 12 2014 fl_transcripts
-rw-rw-r-- 1 htang tigr 125 Oct 14 2014 insertsize
```
Similar Initiatives

OSU Summer Bioinformatics Workshop

- Annual summer workshop started in 2012
- Targeted toward students and faculty with limited background in bioinformatics
- Similar in scope as the JCVI workshop: Instructors present background information, attendees form groups and work together to analyze data and present their findings

- Part of OSU Bioinformatics Graduate Certification program
- Participants learn to use High Performance Computing systems (via OSU HPCC)
- Exposes researchers to iPlant community resources: Atmosphere (cloud), Discovery Environment (workflows)

Peter Hoyt
Dana Brunson
### Similar Initiatives

**OSU Summer Bioinformatics Workshop**

<table>
<thead>
<tr>
<th></th>
<th>Undergrads</th>
<th>Graduate Students</th>
<th>Postdocs</th>
<th>Faculty/staff</th>
<th>Women or Underrepresented groups</th>
<th>Colleges Represented</th>
<th>Universities represented</th>
<th>International Universities</th>
<th>Industries</th>
<th>Govt. Agencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>1</td>
<td>12</td>
<td>2</td>
<td>6</td>
<td>13</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2014</td>
<td>0</td>
<td>20</td>
<td>6</td>
<td>7</td>
<td>27</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>1</td>
<td>32</td>
<td>8</td>
<td>13</td>
<td>40</td>
<td>8</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
Conclusion

• Developed curriculum consisting of diverse topics, maintaining relevance to current advances
• Implemented curriculum as part of training workshops over 4 year period
• Cloud computing technology utilized to expand the reach of the workshop
• Workshop materials made available to the broader community via iPlant
• Teaching material adapted and utilized by similar initiatives
Acknowledgements

JCVI Instructors
• Haibao Tang
• Shelby Bidwell
• Benjamin Rosen
• Maria Kim
• Yongwook Choi
• Agnes Chan
• Christopher Town

JCVI Guest Instructors
• Suman Pakala
• Barbara Methé
• Chuck Merryman

Guest Instructors (US)
• Eric Lyons (Arizona/iPlant)
• Nevin Young (UMN)
• Kevin Silverstein (UMN)
• Andrew Farmer (NCGR)
• Patrick Zhao (Noble Foundation)
• Steven Cannon (USDA-ARS)
• Blake Meyers (DBI)

Guest Instructors (Intl.)
• Heidrun Gundlach (MIPS)
• Jerome Gouzy (INRA)
THANK YOU!