iAnimal

Cyberinfrastructure Enabling Animal Breeding, Genetics, and Genomics

Federate Animal Bioinformatic Platforms with iPlant

PDs

Eric Lyons, Fiona McCarthy, James Koltes, Angus Forbes

Funded by: National Institute of Food and Agriculture/USDA
Rationale

There exists a critical need for easy-to-use computational resources that facilitate management, analysis, and visualization of quantitative data.

Our long-term goal is to develop an iAnimal ecosystem of integrated computational resources to enable agricultural researchers to accelerate their research towards improving US agriculture.

Specific Aims:

- Federating iPlant's cyberinfrastructure with
  - EPIC-CoGe: Genome management, analysis, and comparison
  - AgBase: Functional annotations of agricultural species
  - VCMap: QTL cross species comparisons
- Developing computational systems for
  - Management and visualization of quantitative genomics data
- Developing cyberinfrastructure training resources for
  - 21st century genetics and genomic researchers.
What is iPlant?

Cyberinfrastructure for Life Science Research
What is CoGe?

Accelerating Comparative Genomics

Welcome to CoGe!

Welcome to CoGe, the place for comparative genomics.

CoGe's unique system architecture provides many interconnected tools to create open-ended analysis networks. CoGe also provides an easy-to-use framework to organize and share your data and analyses.

There is no defined place to start an analysis on CoGe, pick any tool below to begin!

CoGe's Tools

- OrganismView: Search for organisms and perform analyses on their genomes
- CoGeBlast: Blast sequences against any number of genomes of your choosing
- FeatView: Search for genomic features by name
- SynMap: Whole genome syntic dotplot analyses
- SynFind: Identify syntenic regions across many genomes
- GEvo: High-resolution sequence analysis of genomic regions
- Latest Genomes

Identifying fractionation and conserved non-coding sequences in maize and sorghum

Questions, problems, suggestions? Contact us
Comparative Genomics with CoGe

Microsynteny

Sequence similarity

Gene model

CoGe link to regenerate the analysis

Gene overlapped by sequence similarity

Syntenic dotplot

Colinear genes

Sequence 1

Sequence 2

Sequence organism: Escherichia coli K12 strain K-12 substrain MG1655 (v2)

http://genomevolution.org/r/6jyz

http://genomevolution.org/r/5yet
What is EPIC-CoGe?

• Adding advanced user-data management
  • Putting you in control of your data

• Integration of functional genomics data
  • Epigenomics, Expression, Genomic Quantitation

• Integration of diversity data
  • Ecotypes, Populations, SNPs

• Integration of JBrowse for Genome Viewing
  • Built-in functionality needed by project
EPIC-CoGe’s user-data management system

Putting you in control of your data:
CoGe is the first read/write genomic platform

• Load your own genome
• Load your own annotations
• Load your own functional genomics data
  • Methylation, Expression, ChIP-Seq, Reads, etc.
• Load your own diversity genomics data (SNPs)
• Metadata and provenance tracking
• History of all run analyses
• Notebooks for organizing datasets
• Keep your data private
• Quickly share with other users
• Mix with public and private data
1. provide annotations (data)
2. provide tools (to help use this data)
3. provide training (to enable use of the data & tools)
1. Functional annotation
   - Gene Ontology
     - 1,612,346 annotations for 314,387 gene products from 74 species
     - molecular interactions (including host-pathogen interactions)

2. Data by Species
   - chicken, turkey
   - horse
   - cow, sheep, pig,
   - catfish, salmon, trout
Comparing QTLs across species
Viewing evolutionary rearrangements across species
Comparing different map types within species
iPlant Federation
Building an ecosystem of interoperability

iAnimal
Shared cyberinfrastructure enables complex workflows

Researcher interested in animal traits

**Step 1. Generate data**
- **RNASeq:** 60 samples from 10 tissues, 2 treatments, 3 replicates

**iPlant Discovery Environment**
- **Step 2:** Transcriptome Assembly
- **Expression Profiling**

**iPlant Data Store**

**Step 3:**
- **VCMap:** Researcher identifies QTLs

**Step 4:**
- **CoGe:** Integrates Expression Profiling, QTL Data, Public Data
  - Identifies 100 genes

**Step 5:**
- **Collaborator:** Examines differentially expressed genes in QTLs

**Step 6:**
- **AgBase:** Functional Annotation of Genes; Identifies Differentially Regulated Pathway

**Collaborator:** Examines differentially expressed genes in QTLs
- Identifies 100 genes

**AgBase:**
- **Step 6:** Functional Annotation of Genes; Identifies Differentially Regulated Pathway

**Shared cyberinfrastructure enables complex workflows**
iPlant Federation

Examples from EPIC-CoGe
Federation
Examples from CoGe
User identity management
iDS data retrieval

Federation

Examples from CoGe

**Load Genome**

**Genome Info**
- Organism: Search
- Version: 
- Type: unmasked: unmasked sequence data
- Source: Search
- Restricted? 
- more...

**Genome Data**
Click the Add Data button to select the data items to be loaded for the genome from: 1) iPlant Data Store, 2) FTP/HTTP site, 3) browser upload, and 4) NCBI. For 1-3, only sequence FASTA files can be loaded at this stage. GFF annotation files can be loaded later on. For 4, both sequence and annotation associated with the specified accession are loaded.

Load Genome

**BETA** Note: this feature is fairly new and still undergoing testing. Please let us know if you experience any problems or have suggestions at coge_genome@gmail.com.
Federation
Examples from CoGe
iPlant Data Store data export; metadata tagging

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoGe Genome ID</td>
<td>9199</td>
</tr>
<tr>
<td>CoGe GenomeInfo Link</td>
<td><a href="http://genomesevolution.org/CoGe/GenomeInfo.pl?gid=9199">http://genomesevolution.org/CoGe/GenomeInfo.pl?gid=9199</a></td>
</tr>
<tr>
<td>CoGe OrganismView Link</td>
<td><a href="http://genomesevolution.org/CoGe/OrganismView.pl?gid=9199">http://genomesevolution.org/CoGe/OrganismView.pl?gid=9199</a></td>
</tr>
<tr>
<td>Genome Name</td>
<td>refgen_v2 assembly (working gene set annotations: 5a)</td>
</tr>
<tr>
<td>Imported From</td>
<td>CoGe: <a href="http://genomesevolution.org">http://genomesevolution.org</a></td>
</tr>
<tr>
<td>ipc-filetype</td>
<td>fasta</td>
</tr>
<tr>
<td>Organism Name</td>
<td>Zea mays (maize; corn)</td>
</tr>
<tr>
<td>Organism Taxonomy</td>
<td>Eukaryota; Viridiplantae; Streptophyta; Streptophytina; Embryop...</td>
</tr>
<tr>
<td>Source</td>
<td>MaizeSequence.org: MaizeSequence.org</td>
</tr>
<tr>
<td>Source Link</td>
<td><a href="http://www.maizesquence.org/index.html">http://www.maizesquence.org/index.html</a></td>
</tr>
<tr>
<td>Type</td>
<td>super masked repeats 50x: repeats with more than 50x occurrence...</td>
</tr>
</tbody>
</table>
RNASeq processing

James Schnable’s qTeller Pipeline

LoadExperiment

NotebookView

Info:
- Name: test
- Description: Experiment
- Type: Experiment
- Restricted: Yes
- Users with access: Coge Team

Contents:
- Experiments (3):
  - @ test (FPKM): Transcript expression measurements (v1, eid2505): Matt Bomhoff
  - @ test (alignment): Mapped reads (v1, eid2503): Matt Bomhoff
  - @ test (read depth): Read depth per position (v1, eid2507): Matt Bomhoff
Details (documentation):

Expression Analysis Pipeline

CoGe can generate gene/transcript expression measurements given a FASTQ input and an annotated genome. Thanks to James Schable, creator of qTeller, for help developing this pipeline!

When a FASTQ file of sequence reads is loaded in LoadExperiment and associated with an annotated genome, the following analysis steps are performed:

1. The FASTQ file is verified for correct format.
2. CutAdapt is run to trim adapter sequence from the reads (parameters: -q 25 --quality-base=64 -m 17).
3. GMAP is run to index the reference genome sequence.
4. GSNAP is run to align the reads to the reference sequence (parameters: --nthreads=32 -n 5 --format=Sam --gmap-mode=none --nofails).
5. SAMtools is run to compute per-position read depth of the resulting alignment (mpileup -D -Q 20).
6. Cufflinks is run to compute per-transcript FPKM (parameters: -p 24).
7. The per-position read depth and per-transcript FPKM values are log transformed and normalized between 0 and 1 for loading.
8. The three results (raw alignment, per-position read depth, and per-transcript FPKM) are loaded as separate Experiments into a Notebook.

Genomes for which this analysis has been performed can have features imported into qTeller. TBD: how to do this …
The AgBase-iPlant Mind Meld...

• iPlant has RNASeq analysis pipelines
  – Sequence assembly, gene annotation, some Diff Expression
• AgBase functional analysis is limited for larger RNASeq data sets
• Provide iPlant capabilities to do genome-wide functional annotation.
  – InterProScan (GO & pathways from sequence)
  – GOanna (GO via Blast searches)
  – Integrating GO annotation data
• Provide tools for GO enrichment analyses of differentially expressed gene sets.
  – Link genomic data with GO annotation data.
  – Tools for functional enrichment
• AgBase Training & Outreach to feature these new pipelines/resources
CoGe is providing support for the Avian Phylogenomics Initiative. This group has sequenced genomes of over 50 avian species, including one species belonging to each major order. The goals of the project are to:

1) Perform phylogenomic analyses to decipher avian phylogeny;
2) Identify genetic changes as signatures of convergent selection, including in vocal learning species;
3) Sequence genomes of avian species that provide scientific value to the community, are national symbols, or endangered; and
4) Provide a public resource for scientists.

The project is partnered with other initiatives such as the Genome 10K (G10K) group to sequencing 10,000 vertebrate species in 10 years, a Copenhagen phylogenome group to decipher avian phylogeny, and the Beijing Genome Institute (BGI) to also decipher avian phylogeny and provide genomic resources. The project is coordinated by Erich Jarvis (Duke), Tom Gilbert (Copenhagen), and Guojie Zhang (BGI).

For select publications from this project, see References.
iAnimal Data Integration

39 Fish Genomes

https://genomevolution.org/CoGe/NotebookView.pl?nid=890
### iAnimal Data Integration

**Cattle, chicken, pig**

#### iAnimal genomes in CoGe

<table>
<thead>
<tr>
<th>Organism</th>
<th>Version</th>
<th>Type</th>
<th>Source</th>
<th>Created</th>
<th>Experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bos taurus (cattle)</td>
<td>UMD3.1</td>
<td>Unmasked</td>
<td>Ensembl</td>
<td>2014-08-06 17:18:24</td>
<td>82</td>
</tr>
<tr>
<td>Bos taurus (cattle)</td>
<td>btau 4.6.1</td>
<td>unmasked</td>
<td>animalgenome.org</td>
<td>2014-08-19 14:15:26</td>
<td>0</td>
</tr>
<tr>
<td>Gallus gallus (Chicken)</td>
<td>4.75</td>
<td>unmasked</td>
<td>Ensembl</td>
<td>2014-08-01 15:11:11</td>
<td>2</td>
</tr>
<tr>
<td>Sus scrofa (pig)</td>
<td>10.2.75</td>
<td>unmasked</td>
<td>Ensembl</td>
<td>2014-08-02 08:39:46</td>
<td>5</td>
</tr>
</tbody>
</table>

#### Mix of public and private data

<table>
<thead>
<tr>
<th>Name of experiment</th>
<th>Type</th>
<th>Source</th>
<th>Groups with access</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bovine Kidney Heart and Pelvic Fat RNAseq</td>
<td>RNAseq</td>
<td>Jim Reesy</td>
<td>Restricted</td>
</tr>
<tr>
<td>Bovine Kidney Heart and Pelvic Fat miRNAseq</td>
<td>RNAseq</td>
<td>Jim Reesy</td>
<td>Restricted</td>
</tr>
<tr>
<td>Bovine Liver miRNAseq</td>
<td>RNAseq</td>
<td>Jim Reesy</td>
<td>Restricted</td>
</tr>
<tr>
<td>Bovine Subcutaneous fat RNAseq</td>
<td>RNAseq</td>
<td>Jim Reesy</td>
<td>Restricted</td>
</tr>
<tr>
<td>Bovine Subcutaneous fat miRNAseq</td>
<td>RNAseq</td>
<td>Jim Reesy</td>
<td>Restricted</td>
</tr>
<tr>
<td>Bovine skeletal muscle miRNAseq</td>
<td>RNAseq</td>
<td>Jim Reesy</td>
<td>Restricted</td>
</tr>
<tr>
<td>Gallus Gallus DNA methylation data</td>
<td>Methylation</td>
<td>Jim Reesy</td>
<td>Restricted</td>
</tr>
<tr>
<td>Gallus Gallus skeletal muscle ChIP-seq</td>
<td>ChIP-seq</td>
<td>Jim Reesy</td>
<td>Restricted</td>
</tr>
<tr>
<td>Sus scrofa adipose RNAseq</td>
<td>RNAseq</td>
<td>Jim Reesy</td>
<td>Restricted</td>
</tr>
<tr>
<td>Sus scrofa iPS ChIP-seq</td>
<td>ChIP-seq</td>
<td>Jim Reesy</td>
<td>Restricted</td>
</tr>
<tr>
<td>Sus scrofa liver RNAseq</td>
<td>RNAseq</td>
<td>Jim Reesy</td>
<td>Restricted</td>
</tr>
<tr>
<td>Sus scrofa muscle RNAseq sample 1</td>
<td>RNAseq</td>
<td>Jim Reesy</td>
<td>Restricted</td>
</tr>
<tr>
<td>Sus scrofa muscle RNAseq sample 2</td>
<td>RNAseq</td>
<td>Jim Reesy</td>
<td>Restricted</td>
</tr>
<tr>
<td>iAnimal Bovine Variants (Notebook with 72 experiments)</td>
<td>SNPs</td>
<td>Jim Reesy</td>
<td>Restricted</td>
</tr>
<tr>
<td>iAnimal Bovine QTL data</td>
<td>QTL data</td>
<td>animalgenome.org</td>
<td>Public</td>
</tr>
</tbody>
</table>

Working with the Community
More Flexible Workflows; New Workflows

In collaboration with James Koltes and Jim Reecy
Want to learn more:
Genome Management and Analysis with CoGe
California Room: Sat. 4-6 pm (today)

Step-by-step tutorial:

• Find and visualize a genome (JBrowse)
• Load your own genome and annotations
• Data management (keeping genomes private and sharing with collaborators)
• Comparative analyses
• Your history
• RNASeq processing

Collaborators: Anyone
We need your feedback and input!

- **Open for business:** happy to help you integrate your genomes, functional genomic datasets, diversity genomic datasets

- **Tell us:**
  - What data you need
  - How you wish to organize, query, and visualize data
    - Advanced analytics
    - Advanced data transformations
    - Advanced data visualizations
Future: Advanced Query, Tool Integration, and Data Visualization
iPlant’s Cyberinfrastructure
The Foundation of the Ecosystem

iPlant Collaborative
iPlant Data Store
iPlant Computational Resources

Cleared Leaf Image Database
Arabidopsis Information Portal

iPlant’s Cyberinfrastructure
The Foundation of the Ecosystem

iPlant Data Store
iPlant Computational Resources

Cleared Leaf Image Database
Arabidopsis Information Portal