AgBioData
Coordinated Innovation Networks Grant

Title: AgBioData: A Coordinated, Collaborative and Innovative Network of Genomic, Genetic and Breeding Databases for Enhanced Agricultural Research Outcomes

USDA
National Institute of Food and Agriculture (NIFA)
Food and Agriculture Cyberinformatics and Tools Initiative (FACT)
https://nifa.usda.gov/program/fact

FACT focuses on data science to enable systems and communities to effectively utilize data, improve resource management, and integrate new technologies and approaches to further U.S. food and agriculture enterprises.

WHAT IS A COORDINATED INNOVATION NETWORK? Coordinated innovation networks projects foster communities that address critical areas by bringing together experts from different disciplines to identify solutions.
Letter of Intent submitted July 25, 2018. Accepted August 15, 2018

Dorrie Main is Project Director. If funded funds will go to:

• Wash State for yearly workshop, and website support (10% FTE)

• Iowa State University for 50% FTE coordinator (J Campbell)

• Phoenix Bioinformatics for Database Sustainability Pilot Study
Objectives:

1: Develop and implement standards for AgBioData data curation

2: Establish common practices for broad use of ontologies, specifically GO, PO, TO and PATO, and provide tools and training for researchers

3: Establish metadata standards across AgBioData members and promote compliance

4: Identify opportunities for a federated model of data exchange for AgBioData member databases.

5: Identify funding options for long term database sustainability

6: Work with funding agencies and journals to enhance data provision by researchers
Timeline:

1. Sept 15: Steering Committee Draft to Whole group
2. Oct 19: Submitted to WSU Grants Office

So, please plan to help in the area most near to your interests between Sept 15 and October 19. THANKS!!!!

We would like to include a representative from each database/resource as an official collaborator on the proposal – will be in touch soon to request letter if this is agreeable

Note about the **Steering Committee**: We are slowly thinking about governance bylaws for AgBioData. We WILL rotate people onto the Steering committee (maybe by election), and we hope to start this process as soon as the grant is in.
TODAYS MEETING:

Genome Nomenclature in various organisms

Ethy/Maggie: Maize
Tanya: Arabidopsis
Sook: Prunus, Cotton
Marcela: Many- Ensemble
Pankaj: Many- Planteome
Taner: Wheat
Nomenclature protocols for maize genomes

Ethalinda Cannon
&
Margaret Woodhouse
MaizeGDB
MaizeGDB holds 10 complete, reference-quality maize genomes with annotation.

- 3 have multiple versions.
- 4 new genomes are nearing completion.
- 25+ new genomes are in progress.

We couldn’t leave naming to the Wild West.

The maize community has a long history of setting, and [mostly] sticking to agreed-upon nomenclature rules. A maize nomenclature committee maintains, and along with MaizeGDB, attempts to enforce the rules.
The state of maize nomenclature

The first 3 B73 reference genomes:

- **V1**: genome = B73 RefGen_v1, AGPv1; annotation = 4a
- **V2**: genome = B73 RefGen_v2, AGPv2; annotation = 5b
- **V3**: genome = B73 RefGen_v3, AGPv3; annotation = 5b+

The genome names were consistent, but the format didn’t allow for genome assemblies for additional maize lines.

The annotations names were unrelated to genome names.

The “+” in “5b+” is a reserved URL character and is still causing occasional problems in MaizeGDB code.

We needed a different nomenclature system.
1. Naming maize genome assemblies

Objectives:
• Unique
• Consistent
• Human- and machine-readable
• Short
• Enforced across the maize research community
• No reserved symbols

This task was surprisingly difficult.
1. Naming maize genome assemblies

Assembly identifiers: (Caution when encoding information in identifiers)

\[ Z(s) - (c) - (D) - (GR) - (V) \]

For instance, B73 version 4 is officially named:

\[ Zm-B73-REFERENCE-GRAMENE-4.0 \]

\[ Zm: Zea mays. \]
\[ B73: The cultivar. \]
\[ REFERENCE: This is the reference genome for B73 (“reference” means a pseudomolecule assembly as opposed to unassembled scaffolds, which are denoted “DRAFT”). \]
\[ GRAMENE: the group that sequenced version 4. \]
\[ 4.0: This is the fourth version of the B73 genome \]
2. Naming maize genome annotations

Annotations have a shorter name.

- Unique across all genome assemblies.
- Identify of genome assembly is built into the name.
- Used to prefix gene model names
- Minimal information encoded in name
2. Naming maize genome annotations

Annotation identifiers:

Z(species)(series)(assembly version)

For instance, For the maize line W22 (the *fourth* genome sequenced since the new naming conventions were established), there are *two* versions of the genome. Therefore:

- The first assembly version: Zm0004a
- The second assembly version: Zm0004b

004 – fourth maize genome assembled  
a/b – version 1 and 2
2. Naming maize genome annotations

Third party annotations

We accept the naming conventions used by third party annotators, for example, GenBank.
3. Naming maize gene models

Objectives:

• Short names.
• Assembly built into name.
• Unique across all genome assemblies.
• Applies only to the “reference” annotation (others use their own conventions, e.g. NCBI).
• Numbering should not imply gene location or order*.

*this is controversial!
3. Naming maize gene models

Gene model names are generated by taking the genome annotation identifier and appending a unique, numerical identifier to it:

\[ Z(\text{species})(\text{series})(\text{version})(\text{random gene identifier}) \]

The unique numerical identifier distinguishes one gene model from another. For example, for B73 version 4:

\[ Zm00001d00001, Zm00001d02002, Zm00001d001224, \text{etc} \]
3. Naming maize gene models

Versioning: distinguishing different annotation versions in the same assembly:

• Assume (hope!) that new annotation versions retain the names of identical gene models.
• Split or merged gene models must get new names.
• To avoid confusion that can arise when there are differences between files downloaded at different times, versions need to be made explicit.
3. Naming maize gene models

Versioning: distinguishing different annotation versions in the same assembly:

\[ Z(\text{species})(\text{series})(\text{assembly version}).(\text{annotation version}) \]

For example, the current annotation for B73 v4 is:

\[ Zm0001d.2 \]

\[ Zm: Zea mays \]

\[ 0001: \text{First genome assembly in series} \]

\[ d: 4^{th} \text{ version of the assembly} \]

\[ 2: \text{second version of the annotation} \]
3. Naming maize gene models

**Versioning:** distinguishing different annotation versions in the *same* assembly:

- Assume (hope!) that new annotation versions retain the names of identical gene models.
- Split or merged gene models must get new names.
- To avoid confusion that can arise when there are differences between files downloaded at different times, versions need to be made explicit.

*The annotation version is not indicated in the gene model name.*
4. Naming pan-genes

A pan-genome is the cumulative diversity within all sequenced maize cultivars in *Zea mays*, including all annotated genes within each cultivar.

Pan-genes are genes that have orthologs in multiple maize cultivars within the pan-genome.
4. Naming pan-genes: ideas

For *curated* genes (genes that have been characterized genetically), pan-genes of curated genes will be given the approved gene symbol (such as *lg1*) in every maize genome where the gene is present.

*human-curated*
4. Naming pan-genes : ideas

For **non-curated* gene models**, all gene models across all maize lines that are in the expected syntenic region will either:

- be given the gene model ID of the first genome sequenced that has the syntelog; or
- be given a new identifier (such as Z0123456) that will be shared as an alias or synonym among all syntelogs (these syntelogs will still retain their unique annotation identifier too)

* Automated process, no human curation
Arabidopsis thaliana genome annotation

The story so far
History of *A. thaliana* genome annotation

- Original annotation:
  - 2000: Completion of genome sequence, TIGR1 genome release

- Reannotation (10 versions)
  - 2005: TAIR6 genome release
  - 2016: Araport11 genome release

- 20??: Next genome release
Source Germplasm = Col-0

- **Past:** Multiple Col-0 stocks used, unclear how these were related to each other.
- **Proposal:** Col-0 seed stock CS70000 designated as the reference seed stock.
Naming Conventions

- AT[1-5,C,M]gNNNNNN
- Locus: At1g01020
- Gene model: At1g01020.1, At1g01020.2, ...
• Currently no distinction in naming between genome assembly (pseudo-chromosomes) and gene annotation (individual gene calls)
  – TAIR10 genome assembly (same as TAIR9)
  – TAIR10 gene annotation (different from TAIR9)
• Not ideal approach – need to be able to version/name them independently
Existing/Upcoming issues

- Ler sequence published—used AGI identifier approach but inconsistent use between Col-0 and Ler use
  - Col-0 At1g01040 != Ler At1g01040
  - Created mapping file with author-supplied source files
- ‘Platinum’ standard, PacBio sequences (de novo assembly) of Col-0, Ler, up to 50 other ecotypes coming—need for a consistent nomenclature approach
- Many other ecotypes are sequenced with Col-0 reference guided assembly (1001 genomes project) and could/should be annotated
- Variation in gene insertion/deletion/rearrangement/modification in other ecotypes vs. Col-0
Handling user feedback

• Accumulate suggestions between releases
  – Need for tracking, verification, standards for acceptance of edits

• Incremental updates? How to propagate these efficiently and effectively – NCBI/DDBJ/EMBL and others
GENOME NAMING GUIDELINE IN GDR
(AND COTTONGEN, CGD, CSFL)

Sook Jung
Washington State University
Genome Naming Guideline

- [Genus] [species] genome v[assembly version].a[annotation-version]
- Example: Prunus persica genome v2.0.a1
- Where:
  - Genus = the genus of the organism
  - Species = the species of the organism
  - Assembly version = the version of the assembly with a major and minor number. The major version is incremented with major changes or releases of the assembly and the minor number is incremented when minor changes are made to the assembly.
  - Annotation-version = a single numeric value that is incremented each time a new annotation is released. It restarts at 1 each time the assembly version is incremented.
It works most of the time
(GDR has 21 genome assemblies from 7 crops and 14 species)

- Prunus persica Genome v2.0.a1
- Prunus persica Genome v1.0
- Prunus avium Genome v1.0.a1
- Pyrus communis Genome v1.0
Multiple genome assemblies from the same species – add accession name

- Malus x domestica Genome v1.0.a1
- Malus x domestica Genome v2.0.a1
- Malus x domestica Genome v3.0.a1
- Malus x domestica GDDH13 Whole Genome v1.1
  - Up to v3.0 was done using heterozygous Golden Delicious, and the newest version was done using GDDH13, a doubled-haploid Golden Delicious tree.
Multiple genome assembly with the same accession? – add institution name.

◦ CottonGEN example
  ◦ Gossypium hirsutum (AD1) acc 'TM-1' genome CGP-BGI v1.1 assembly & v1.0 annotation
  ◦ Gossypium hirsutum (AD1) acc 'TM-1' genome NAU-NBI v1.1 assembly & v1.1 annotation
  ◦ Gossypium hirsutum (AD1) acc 'TM-1' genome UTX-JGI v1.0 assembly v1.0

◦ Same accession and institution? Use published name.
  ◦ Rosa chinensis Genome v1.0
  ◦ Rosa chinensis Old Blush homozygous Genome v2.0
    ◦ Rosa chinensis Old Blush Illumina genome v1.0
Sometimes authors skip versions when they publish:

- Fragaria vesca Genome v1.0.a1
- Fragaria vesca Genome v1.1.a1
- Fragaria vesca Genome v1.1.a2
- Fragaria vesca Genome v2.0.a1
- Fragaria vesca Genome v2.0.a2
- Fragaria vesca Genome v4.0.a1
Some times authors publish duplicated names

- Fragaria vesca Genome v1.0.a1
- Fragaria vesca Genome v1.1.a1
- Fragaria vesca Genome v1.1.a2
- Fragaria vesca Genome v2.0.a1
  - Fragaria vesca Genome v1.1.a2 (Darwish et al. 2014) track – we originally called it Fragaria vesca Genome v2.0.a2
- Fragaria vesca Genome v2.0.a2
  - Fragaria v2.0.a2 (Li et al. 2017)
conclusion

- Need to refine the naming system
- Need to work with journals and communities
Perspectives on genome nomenclature from 56 species

Marcela Karey Tello-Ruiz, PhD

September 5th, 2018

For AgBioData Consortium Nomenclature Discussion
Various Gene IDs systems @ Gramene/Ensembl Plants

- Reuse gene IDs assigned by the genome projects responsible for the annotation and provided to the INSDC:
  - NCBI: GenBank
  - EMBL-EBI: ENA
  - DDBJ: DDBJ
- Gramene/Plant Reactome/Ensembl Plants work with the INSDC (mainly ENA) to resolve IDs (prioritizing UniProt proteome-based IDs, if applicable).
- Adopt a system agreed/adopted by the community early on
- Provide a mechanism (ID converter) for mapping back to IDs in previous versions, when possible
- MODs and other resources work with GenBank/ENA to get a clear picture of how IDs are constructed
<table>
<thead>
<tr>
<th>Species</th>
<th>Assembly</th>
<th>Gene Annotation</th>
<th>Gene ID example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis lyrata</td>
<td>Araly1.0</td>
<td>Araly1.0</td>
<td>Al_scaffold_0001_1000</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>TAIR10</td>
<td>AraPort11</td>
<td>AT3G52430</td>
</tr>
<tr>
<td>Chlamydomonas reinhardtii</td>
<td>v5.5 (GCA_000002595.3)</td>
<td>JGI via ENA</td>
<td>CHLRE_15g637761v5</td>
</tr>
<tr>
<td>Oryza sativa japonica</td>
<td>IRGSP-1.0</td>
<td>RAP-DB</td>
<td>Os05g0113900</td>
</tr>
<tr>
<td>Solanum lycopersicum</td>
<td>SL2.50</td>
<td>ITAG2.3</td>
<td>Solyc01g087250.2</td>
</tr>
<tr>
<td>Sorghum bicolor</td>
<td>V3 (GCA_000003195.3)</td>
<td>JGI via ENA</td>
<td>SORBI_3004G141800</td>
</tr>
<tr>
<td>Triticum aestivum</td>
<td>IWGSC v1.0</td>
<td>IWGSC v1.0</td>
<td>TraesCS3D01G273600</td>
</tr>
<tr>
<td>Vitis vinifera</td>
<td>IGGP 12x</td>
<td>2012-07-CRIBI</td>
<td>VIT_01s0010g03900</td>
</tr>
<tr>
<td>Zea mays</td>
<td>B73_RefGen_v4</td>
<td>MAKER-CSHL</td>
<td>Zm00001d048577</td>
</tr>
</tbody>
</table>

- In the near future, we will see the assemblies stabilizing, but the annotations updated based on evidence.
- Gene annotations: protein-coding, non-coding RNAs, pseudogenes.
- Regulatory features???

Full table in Gramene’s release notes (56 species) [http://www.gramene.org/release-notes-58](http://www.gramene.org/release-notes-58)
Challenges

1. Lack of continuity in gene IDs/names
2. Caution on including too much information in a name
3. Propagating gene information between versions of the reference assembly. Many different approaches based on the quality of the draft assembly and the complexity of the organism.
4. Challenges associated with initial submission and updates to NCBI
5. Not all proteomes fully represented/updated @ UniProt
6. PanGenomes: Moving from a single reference to many references
   • Same as above but more complex
   • Learn from experience: *Arabidopsis thaliana* & *A. lyrata*
   • Grape: suggestion to include species and cultivar names (prefix/suffix letter codes) and numeric ID across species (some gene models will be species-specific).
Proposed naming strategy for the grape pangenome

- Annotate species and varieties in a gene ID.
  - *Species*: Community follows UniProt recommendation. First two letters of species name as prefix. Example: **Vitci** for *Vitis cinerea* (ask UniProt to add species if not in the list: [http://www.uniprot.org/docs/specieslist](http://www.uniprot.org/docs/specieslist)).
  - *Cultivar*: Annotate the cultivar in the suffix (2-4 letters code), as they are basically considered different alleles. Examples: Use the prefix **Vitvi** (*V. vinifera*) for cabernet sauvignon (cs), pinot noir (pn), and flame, then suffix for cultivar (**Vitvi00g0000-cs** is a cab gene). Concord is a *V. labrusca* (**Vitla**), add suffix.
- Relating orthologous genes. Compare the sequences to the latest release of a gene annotation for IDs to match. A challenging workflow that requires renaming and transformation of the IDs...
Convert assembly-specific & cross-reference gene IDs
Suggestions towards standardizing gene nomenclature

• Once the data is released from INSDC, everyone must use the same ID
• Prioritize representation of (up-to-date) proteomes in UniProt
• Gene IDs independent of assembly version, genomic location, function & orthology
• Adopt standards from good quality nomenclature systems (e.g., human, yeast, Arabidopsis)
Suggestions towards standardizing gene nomenclature

• Standard tools to convert IDs across resources, assemblies, etc.

• **Community collaborative platform** - Build a 'data wiki' and let communities provide their own names (Dan Bolser)
Gene Names: Ensembl gene name projection pipeline

- Up to 90% of gene names could be projected from closely related species (e.g. between the rice).
- For which pairs of species does it make sense to project genes between?
  - E.g. over what taxonomic range and with what stringency?
Gene Nomenclature
Planteome Strategy

Pankaj Jaiswal
Oregon State University

September 5, 2018
AgBioData Group Meeting
Genomes

• Reference genomes:
  • Accession/germplasm/ecotype-specific

• Non-reference genomes
  • For genetic diversity (Mainly for SNPs but if fully sequenced can be used for finding new and novel genes)
    • Accession/germplasm/ecotype-specific

• Pan-genomes
  • Species-level
  • Genus-level
Gene set changes observed

- Identify new gene (includes coding and noncoding)
- Identifying null allele (gene missing in an ecotype/accession)
- Existing genes
  - Start and stop coordinates may change
  - New/altered UTRs
  - Different transcript isoforms
    - Novel splicing
    - New transcribed region
    - New peptide
  - Mergers
  - Splits (may need adding a minimum of one new gene)
  - Obsolete/delete
Some points to consider

• Need consistency for the community and semantic querying, NLP and data updates
• Keep the Super gene set at the Species/genus level considering we will see a lot of Pan-genome projects and germplasm-specific gene sets
• Populate the super gene set by adding uniques to the pool
• For each version of the annotation (not assembly) map to the super pool and borrow the IDs or create a new one for new genes
• Never create version dependent IDs or insert version# in gene ID
• Make an arrangement with INSDC/GenBank to deposit the new annotation (proteomes etc)—THEY WILL TAKE CARE OF THE MAPPING AND VERSION NUMBER as long as IDs are consistent.
  • MANY genomes lack/do not deposit annotations
Some points to consider

• If the users perform structural annotation to revise the existing model, continue using the same gene ID at the locus
  • Release it with the new annotation version periodically
  • Second/third party annotations: They need to map to the super set IDs and share their data to become part of the official release. EVERYONE Needs to use the same common set of updated gene sets.

• If a new gene is added, create a new ID
  • If the genome assembly is in the pseudomolecule form use the number series and zero padded 100/1000\textsuperscript{th} space to maintain series
  • If the genome is in scaffolds use the next available number
  • Release it with the new annotation version periodically (quarterly/bi/annual)